Epidemiology & Health Data Insights (eISSN 3080-8111)

Editorial. *Epidemiol Health Data Insights. 2025;1(1):ehdi002.* https://doi.org/10.63946/ehdi/16387

Where Epidemiology Meets Data Science: The Dawn of a New Platform Bridging Evidence, Analytics, and Policy

Praveen Kumar Potukuchi [1]

[1] Principal Health Economist, Boston Scientific Corporation, 300 Boston Scientific Way, Marlborough, MA, 01752, USA.

Received: May 15, 2025
Accepted: May 16, 2025

Corresponding author: Praveen K Potukuchi

Email: praveen.potukuchi@gmail.com

This work is licensed under a **Creative Commons Attribution 4.0 International License**

Epidemiology & Health Data Insights is a new brand international, open-access, peer-reviewed journal dedicated to advancing the science and practice of epidemiology through the integration of health data science. Our mission is to bridge traditional epidemiological methods with modern analytical approaches to address pressing global health challenges. By providing a platform for multidisciplinary collaboration, the journal supports evidence-based solutions that improve public health outcomes, promote equity, and guide healthcare policies.

It is with great pleasure that we present the first issue of **Epidemiology & Health Data Insights**, an international, open-access, peer-reviewed journal committed to advancing the field of epidemiology through the integration of health data science. This journal is launched in response to the growing need for multidisciplinary approaches to address global health challenges with evidence-based, data-driven insights.

Our mission is to foster collaboration among epidemiologists, biostatisticians, data scientists, public health professionals, and policymakers to generate actionable knowledge that informs health equity, policy, and practice worldwide. Topics of focus include disease surveillance, predictive modeling, big data applications, real-world evidence, and public health interventions.

We are proud to announce that our editorial process will be guided by an international team of highly proficient and accomplished experts in epidemiology, biostatistics, data science, and public health. Their collective expertise ensures a rigorous, transparent, and fair peer-review process, maintaining the highest standards of academic integrity. We warmly invite researchers, practitioners, and policymakers to submit their original research, reviews, case studies, and methodological innovations. Together, we can expand the boundaries of public health knowledge and practice.

Welcome to **Epidemiology & Health Data Insights**. We look forward to your contributions and to building a vibrant community of scholars committed to advancing global health through data-driven science.

Epidemiology & Health Data Insights (eISSN 3080-8111)

Review Article. *Epidemiol Health Data Insights. 2025;1(1):ehdi004* https://doi.org/10.63946/ehdi/16385

National Electronic Oncology Registry in Kazakhstan: Patient's Journey

Altynay Beyembetova[1], Ayana Ablayeva[1], Ruslan Akhmedullin[1], Diyora Abdukhakimova[1], Aigerim Biniyazova[1], Abduzhappar Gaipov[1]

[1] Department of Medicine, Nazarbayev University School of Medicine, Astana, Kazakhstan

Received: May 05, 2025
Accepted: May 16, 2025

Corresponding author:

Altynay Beyembetova

Email
altynay.akhmedinova@nu.edu.kz

This work is licensed under a **Creative Commons Attribution 4.0 International License**

Abstract: The burden of non-communicable diseases (NCDs), particularly cancer, is steadily increasing in Kazakhstan alongside socioeconomic development, highlighting the urgent need for effective national management and control strategies. The Electronic Registry of Oncological Patients (EROP) is a key initiative designed to support this goal by enabling comprehensive surveillance of cancer patients across the country. This review examines the patient journey that informs EROP's data collection process and explores its potential as a foundation for local cancer epidemiology research. Analysis of the registry reveals extensive coverage of dispensary, outpatient, and inpatient visits, capturing detailed socio-demographic information, diagnoses, treatments, and mortality outcomes. EROP provides valuable insights into cancer morbidity, mortality, and quality of life across spatial and temporal dimensions in Kazakhstan. The review also addresses current limitations and underexplored aspects of the registry, emphasizing the importance of careful interpretation to ensure robust, evidence-based research.

Keywords: Oncology; Cancer; National registry; Epidemiology;

Introduction

Over the past three decades, global health has shifted from being dominated by communicable diseases to facing a rising burden of non-communicable diseases (NCDs). This shift was primarily driven by factors such as urbanization, aging populations, lifestyle changes, and improvements in infectious disease control, which have been especially pronounced in low- and middle-income countries (LMICs). According to the World Health Organization (WHO), NCDs account for over 70% of global deaths, with a significant proportion occurring in developing regions [1]. In Central Asia, this transition mirrors global trends but is shaped by unique regional dynamics. Following the dissolution of the Soviet Union in 1991, Central Asian countries faced economic instability, healthcare system disruptions, and a legacy of underinvestment in chronic disease management [2]. During the 1990s, infectious diseases remained a primary concern in the region [3], which remained so in the following 30 years, thanks to advancements in health policy, which made CDs less prominent. By the 2010s, cancer had become a significant public health challenge in the region, reporting increasing incidence rates linked to smoking, dietary shifts, and environmental exposures [4]. This epidemiological transition necessitated an appropriate reorientation of health systems to resolve emerging challenges of chronic conditions like cancer.

In Kazakhstan, the transition from infectious diseases to non-communicable diseases (NCDs) has been particularly notable. In the post-Soviet era, the country grappled with infectious disease outbreaks and a fragmented healthcare infrastructure [5]. However, beginning in the early 2000s, economic growth — largely driven by the energy sector — enabled substantial improvements in healthcare infrastructure and access [6]. Paradoxically, this socioeconomic progress also facilitated behavioural and environmental changes that elevated NCD risks, including increased alcohol and tobacco consumption, sedentary lifestyles, and dietary shifts [1]. Lung, stomach, and colorectal cancers are among the most prevalent malignancies, influenced by both genetic predispositions and modifiable risk factors [7].

Considering these developments, robust data systems became essential for understanding cancer trends, informing policy strategies, and shaping evidence-based interventions. The Electronic Registry of Oncological Patients (EROP) represents a critical step in Kazakhstan's efforts to modernize cancer surveillance and research.

This review aims to explore the research potential of EROP, emphasizing its value as a foundational data source for advancing cancer epidemiology, monitoring clinical outcomes, and supporting strategic health policy planning in Kazakhstan and comparable LMIC contexts. By examining the structure, content, and analytical possibilities offered by EROP, this work seeks to underscore its significance in the broader landscape of digital health transformation and non-communicable disease control.

Registry forms

The Electronic Registry of Oncological Patients (EROP) contains collective data on oncology patients across all regions of Kazakhstan from 2014 to 2023. This information has been gathered from four standardized forms used for documenting and tracking medical information and compiled throughout an oncological patient's journey (Figure 1).

Form 034/u - "Notice": When a patient is first diagnosed with an oncological disease or identified as at risk of developing one, Form 034/u is filled out to document this initial notice.

Form 030-6/u - "Control Card of Dispensary Observation (Onco)": Following diagnosis, the patient is registered at the local oncological dispensary for monitoring and surveillance. At this stage, Form 030-6/u is completed for each patient, including details such as registration and deregistration dates.

Form 012/u - "Statistical Card of a Patient Discharged from Hospital": If the patient undergoes hospitalization, whether planned or urgent, Form 012/u is finalized and uploaded upon discharge. This form records data for all oncology inpatients, covering both 24-hour and daytime stays and including information on the diagnostic and therapeutic services received during the hospitalization.

Form 052/u - "Medical Record of an Outpatient": Information about ambulatory oncology patients in the registry is sourced from Form 052/u, which documents the details of outpatient care. The form records data on ambulatory services patients received.

For each form, the data is entered by attending physicians during respective patient encounters through authorized medical portals called healthcare or medical information systems (HIS or MIS) [8]. As part of UNEHS, the recorded forms from medical organizations all over the country are then accumulated at the Republican Center for Electronic Health (RCEH) for healthcare statistics. For epidemiological studies, raw data separately provided for each form is securely transferred to the local servers of a research facility, such as a university, with access strictly limited to authorized personnel [8]. For this review, the end data of EROP collected throughout this patient journey were grouped and discussed for potential use in healthcare research. The study was approved by the Institutional Review Board of Nazarbayev University (protocol code: NU-IREC 651/24112022 and date of approval: 28 November 2022), with exemption from informed consent.

Registry description

The data provided on EROP

The information provided on EROP spans from simple demographic information to details of cancer progression and treatment received at medical facilities. The data variables available in EROP are provided in Table 1 and grouped into seven categories. .

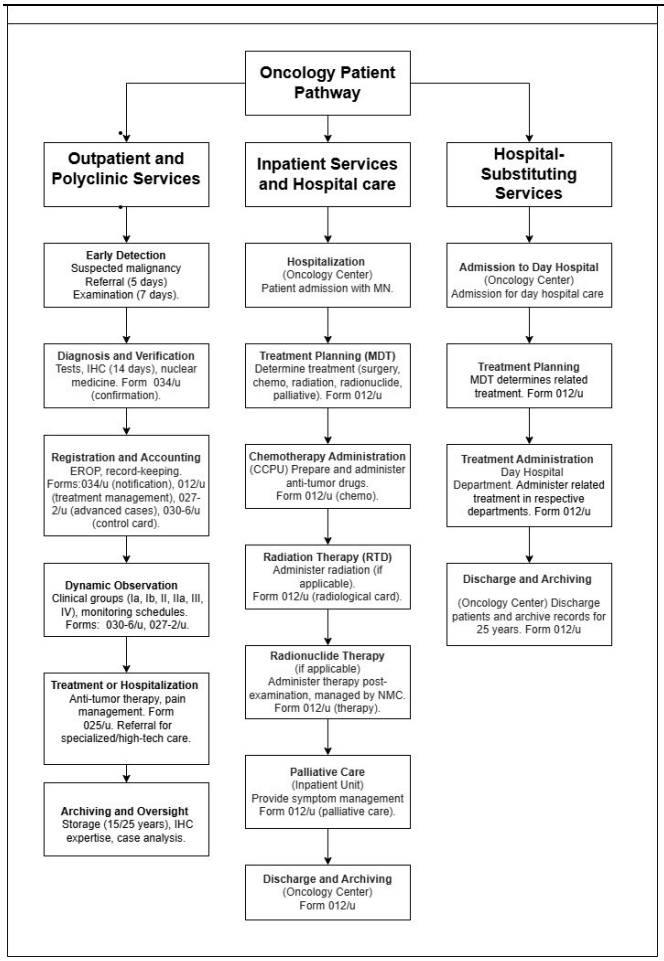

Oncology Patient Pathway
Outpatient and Polyclinic Services
Inpatient Services and Hospital care
Hospital-Substituting Services
Early Detection
Suspected malignancy
Referral (5 days)
Examination (7 days).
Diagnosis and Verification
Tests, IHC (14 days), nuclear medicine. Form 034/u (confirmation).
Registration and Accounting
EROP, record-keeping. Forms:034/u (notification), 012/u (treatment management), 027-2/u (advanced cases), 030-6/u (control card).
Dynamic Observation
Clinical groups (Ia, Ib, II, IIa, III, IV), monitoring schedules. Forms: 030-6/u, 027-2/u.
Treatment or Hospitalization
Anti-tumor therapy, pain management. Form 025/u. Referral for specialized/high-tech care.
Archiving and Oversight
Storage (15/25 years), IHC expertise, case analysis.
Hospitalization
(Oncology Center)
Patient admission with MN.
Treatment Planning (MDT)
Determine treatment (surgery, chemo, radiation, radionuclide, palliative). Form 012/u
Chemotherapy Administration
(CCPU) Prepare and administer anti-tumor drugs. Form 012/u (chemo).
Radiation Therapy (RTD)
Administer radiation (if applicable).
Form 012/u (radiological card).
Radionuclide Therapy
(if applicable)
Administer therapy post-examination, managed by NMC. Form 012/u (therapy).
Palliative Care
(Inpatient Unit)
Provide symptom management
Form 012/u (palliative care).
Discharge and Archiving
(Oncology Center)
Form 012/u
Admission to Day Hospital
(Oncology Center)
Admission for day hospital care
Treatment Planning
MDT determines related treatment. Form 012/u
Treatment Administration
Day Hospital Department. Administer related treatment in respective departments. Form 012/u
Discharge and Archiving
(Oncology Center) Discharge patients and archive records for 25 years. Form 012/u

Figure 1. Oncological Patient Journey in EROP.

Identification numbers

Each patient in UNEHS, including EROP, is assigned with unique identification number, called RpnID. RpnID helps to connect information of individual patients from different registries of UNEHS without using patient-identifiable information and compromising the confidentiality of patients. This variable also helps to connect EROP data with other databases of the UNEHS. For example, in extracting information on the death status of the patient or comorbidities he or she has, alongside the main disease of interest. For example, in extracting information on the death status of the patient or comorbidities he or she has alongside with the main disease of the interest.

Provider information

Provider information as an oncological dispensary, where oncological patients were assigned for surveillance or treatment, as well as the medical clinic providing primary medical care, is available in EROP. These variables help to assess the oncological service across regions of Kazakhstan and to evaluate the regional burden of disease.

Demographic information

Demographic information in EROP includes birthdate, sex, ethnicity, residency, and region. The birth date variable given in Day/Month/Year format helps to calculate the age of patients at the moment of diagnosis, treatment and death. The derived age variables can help to see the trend of morbidity, mortality and quality of life estimates for oncological patients across age strata. Information on sex and ethnicity aids in stratifying the burden of oncological diseases across different demographic categories. Certain forms in the EROP include details about the patient's living region, based on the administrative structure of Kazakhstan for the corresponding year, encompassing oblasts and administrative cities. Furthermore, residency variables provide additional insights into whether patients reside in urban or rural areas.

Diagnoses

Each diagnosis in EROP forms is recorded using International Classification of Diseases 10th Revision (ICD-10) codes, enabling the extraction of specific oncological diseases from the database. Additionally, certain EROP forms include diagnosis data based on the International Classification of Diseases for Oncology (ICD-O) codes, which provide detailed information on the morphology and topology of oncological diseases. Further diagnostic details are captured using variables for staging, based on the TNM system. This includes information on the stage of the disease, provided separately, along with details on tumor (T), lymph nodes (N), and metastasis (M). The date of diagnosis is recorded in Day/Month/Year format, allowing for the analysis of temporal trends in disease burden. Some forms also contain information about the circumstances under which the disease was identified, facilitating the evaluation of screening effectiveness, where it is applicable.

Stay in medical organization

The 012/u form of EROB provides comprehensive details about hospital stays. It includes admission and discharge dates formatted as Day/Month/Year, alongside information on the unit and bed profile associated with each hospitalization and discharge event. Furthermore, it specifies whether the hospitalization was planned or urgent, and provides information about the purpose of the hospitalization. This information offers significant insights into the burden that oncological diseases place on healthcare facilities. They also aid in analyzing patient flow within hospitals, contributing to a better understanding of operational dynamics and resource allocation.

Table 1. EROP framework.

Data elements and variables	Description	Potential application in research	Presence in EROP forms
Identification numbers			
RpnID	Unique identification number assigned to every patient in UNEHS	Interconnection of UNEHS databases	012/u, 030-6/u, 052/u, 034/u
Provider information			
Oncological dispensary	One of 24 facilities in Kazakhstan where patients are registered for oncological care and surveillance	Region identification and assessment of regional burden	012/u, 030-6/u, 052/u
Registered clinic	Clinic where patients are registered and get primary medical care	Region identification and assessment of regional burden	012/u, 030-6/u
Demographic information			
Birth date	Recorded date of birth in format of Day/Month/Year	Calculation of age at diagnosis, age at death and age-specific estimates of burden of disease	012/u, 030-6/u
Sex	Male/Female	Sex-specific estimates of burden of disease	012/u
Ethnicity	Ethnic identity like Kazakh, Russian and others	Ethnicity-specific estimates of burden of disease	
Region	Administrative region of Kazakhstan, where patient is from	Regional burden of disease	012/u
Residency	Type of residence, such as urban (republican or regional cities) or rural (villages and small towns)	Residency-specific estimates of burden of disease	012/u
Diagnoses			
ICD-10 code	ICD-10 code for main oncological disease and comorbidities	Extraction of specific oncological diseases from database	012/u, 030-6/u, 052/u
ICD-10 code name	Description of diagnosis	Distinction of types within oncological diseases	

Diagnosis date	Recorded date of diagnosis in format of Day/Month/Year	Identification of temporal estimates of disease incidence	012/u
Cancer stage	The level of cancer progression, classified using the TNM system, from stages I to IV	Stage-specific stratification of oncological disease cases	012/u, 030-6/u, 034/u
T	Tumor size	Analysis of progression and treatment outcomes	012/u, 030-6/u
N	Number of affected lymph nodes	Tracking of disease spread and staging	012/u, 030-6/u
M	Presence of metastasis	Study of dissemination and impact on survival	012/u, 030-6/u
ICD-O morphology code	Code representing the histological structure and cellular characteristics of the oncological disease	Classification of oncological diseases based on histological characteristics	030-6/u
ICD-O topology code	Code indicating the anatomical site of the oncological disease	Classification of oncological diseases based on their anatomical locations	030-6/u
Diagnosis identification	How a disease was diagnosed, such as during screening, check-up, or self-initiated consultation	Understanding of patient pathways to diagnosis and the impact on treatment outcomes	030-6/u
Stay in medical organization			
Hospital admission date	Recorded date of hospitalization in format of Day/Month/Year	Identification of unique hospital stay cases	012/u
Hospital discharge date	Recorded date of hospital discharge in format of Day/Month/Year	Identification of unique hospital stay cases	012/u
Type of hospitalization	Planned/Urgent	Comparison of planned and urgent care usage, assessment of emergency and preventive care systems	012/u
Purpose of hospitalization	Reasons for hospital stay, such as examination, treatment, continuation of care, or rehabilitation	Insights into reasons for hospital stays supporting analysis of treatment trends and resource use	012/u
Admission unit	Hospital unit where the patient was admitted	Usage patterns of oncological care units	012/u
Discharge unit	Hospital unit from which the patient was discharged	Patient flow in oncological care units	012/u
Admission bed profile	Assigned bed at admission	Bed allocation for oncological care	012/u
Discharge bed profile	Bed assigned before discharge	Bed usage and turnover in care units	012/u
Therapy			
Type of therapy	Types of therapy as chemotherapy, radiotherapy, surgery, and immunotherapy	Analysis of effect of therapy types on disease outcomes	012/u

Type of chemotherapy	Type of chemotherapy as adjuvant (preventive, therapeutic), neoadjuvant, independent	Analysis of effect of chemotherapy types on disease outcomes	012/u
Side effects of chemotherapy	Side effects as gastrointestinal, hematological, hepatitis, infectious, other or absent	Analysis of side effects of chemotherapy	012/u
Type of radiotherapy	Types of radiotherapy as beta therapy, gamma therapy, combined radiation therapy, photon radiation therapy	Analysis of effect of radiotherapy types on disease outcomes	012/u
Complications of radiotherapy	Complications as local, on regional lymph nodes, systemic, without	Analysis of side effects of radiotherapy	012/u
ICD 9 code for surgical procedures	Code indicating performed surgical procedures	Analysis of surgical procedures on disease outcomes	012/u
Chemotherapy start date	Recorded date for the initiation of chemotherapy in format of Day/Month/Year	Calculation of chemotherapy duration	012/u
Radiotherapy start date	Recorded date for the initiation of radiotherapy in format of Day/Month/Year	Calculation of radiotherapy duration	012/u
Medical procedure date	Recorded date for the performance of medical procedure in format of Day/Month/Year	Calculation of days after surgical procedures	012/u
Outcomes			
Death date	Recorded date of death in format of Day/Month/Year	Survival analysis of oncological patients	030-6/u
Outcome of hospital stay	Outcome of hospital stay like discharge, death, transfer	Insight into hospital performance and patient survival rates	012/u
Treatment outcome	Outcome of received care during hospital stay as improvement, no change, deterioration, death	Evaluation of care effectiveness and disease progression	012/u
Financial costs			
Cost of medical services	Total cost of services provided per one hospital stay or ambulatory visit in KZT	Analysis of economic burden of oncological diseases	012/u, 052/u
Funding source	Sources of funding like social health insurance, out-of-pocket expenses, and sponsorship contributions	Evaluation of financial resource distribution and utilization	012/u

Therapy

The database includes detailed information on therapies administered during hospitalization. Specific details are provided for chemotherapy, radiotherapy, and surgical procedures. For chemotherapy, data includes the type, side effects, and start date (Day/Month/Year format). Radiotherapy information covers the start date, type, and associated complications. Surgical procedures are recorded with ICD-9 codes and performance dates. Additionally, data on medications administered as part of chemotherapy or immunotherapy are available. This comprehensive dataset facilitates the analysis of treatment patterns, effectiveness, and outcomes in oncological care.

Outcomes

Several outcomes of oncological diseases are documented in EROP, tailored to the type of form used. Treatment outcomes include improvement, no change, disorientation, or death for patients who received therapy. Hospital outcomes capture processes related to hospital stays, such as discharge, continuation, transfer, or death. Additionally, the death date (Day/Month/Year format) is recorded for patients whose death was registered. These variables enable analysis of treatment effectiveness, patient outcomes, and healthcare processes.

Financial costs

Financial costs for services per hospital stay and an ambulatory visit are provided in KZT. Additionally, the funding sources for hospitalization services are detailed. This data facilitates the analysis of the economic burden of oncological diseases across various social demographic groups, spatial regions, and temporal periods.

The data provided on EROP

The data from EROP are accessed with permission of the study from the Institutional Review Board and a contractual agreement with the provider. The database does not contain identifiable elements as the name and national ID of the patient; instead, unique population IDs called RpnID used in the medical systems are available for data merge and connection. For additional protection of data, on the level of the data receiver, databases are uploaded and only accessed on restricted electronic servers, with access given to only authorized users.

Discussion

The Registry serves as a comprehensive national oncology database. With near-complete coverage and systematic reporting from oncology centres across the country, EROP enables robust epidemiological surveillance and supports longitudinal and retrospective analyses. Its structure facilitates disaggregation by tumour site, age, sex, and geographic region, allowing researchers to assess patterns in cancer incidence, mortality, treatment, and survival.

Several studies have utilized EROP to examine cancer epidemiology, including incidence and diagnostic characteristics. Shatkovskaya et al. (2021) analyzed data from 2014 to 2019 to explore the local epidemiology of HR+/HER2– advanced breast cancer, focusing on stage diagnosis and treatment patterns. In a more comprehensive national study, Midlenko et al. [9] assessed breast cancer epidemiology among women aged 25 years and older, stratifying the findings by age group and geographic region. Pediatric oncology has also been investigated through EROP: Mussina et al. [10] conducted an age-specific cohort study of hematological malignancies in children under 18 between 2014 and 2021, confirming the registry's capacity for pediatric cancer surveillance. In the field of gynecologic oncology, Sakko et al. [11] used registry data to examine the prevalence and temporal dynamics of major cancer-related surgeries, such as hysterectomies and adnexal procedures, linking surgical trends to diagnostic patterns.

In addition to incidence and prevalence, EROP has been instrumental in supporting analyses of survival and mortality. Its integration with death certificate data through the Patient Registry and hospital discharge records enables the estimation of in-hospital, 30-day, 1-year, and 5-year mortality. Studies by Midlenko et al. [9] and Mussina et al. [10] applied Cox proportional hazards regression to identify clinical and sociodemographic predictors of mortality. Using a linked registry approach, Shatkovskaya et al. [12] estimated five-year overall survival in patients with metastatic urothelial cancer diagnosed in 2017, demonstrating the registry's applicability for long-term outcome evaluation. Similarly, Umurzakov et al. [13] investigated incidence, mortality, and survival among male reproductive cancers at the subnational level, highlighting the value of EROP in regional comparative studies.

Akhmedullin et al. [7] conducted a national trend analysis of cancer-related mortality between 2014 and 2022, employing Jointpoint regression models to calculate Average Annual Percent Change (AAPC), as well as estimating mortality-to-incidence ratios (MIR) and proportionate mortality (PM). Their findings offer valuable insights into the evolving national cancer burden and the effectiveness of health policies over time.

Beyond epidemiological indicators, EROP has supported investigations into real-world treatment patterns and health system performance. Shatkovskaya et al. [14] analyzed the use of endocrine and targeted therapies in patients with HR+/HER2– metastatic breast cancer, documenting systemic therapy uptake and clinical decision-making. In gynecologic oncology, Sakko et al. [11] assessed operative trends and regional variations, identifying deviations from international recommendations and proposing updates to national clinical guidelines. Furthermore, EROP has been used to evaluate diagnostic delays, stage at presentation, and time to treatment initiation—indicators critical for assessing system responsiveness.

Finally, EROP facilitates the analysis of geographic disparities and health equity. Disaggregated data allow for comparisons across urban and rural settings, highlighting inequalities in cancer detection, treatment access, and outcomes. Studies by Midlenko et al. [9], Umurzakov et al. [13], and Akhmedullin et al. [7] collectively demonstrate the utility of the registry for identifying regional resource gaps, monitoring temporal trends, and informing national and subnational cancer control strategies.

The strengths and limitations of the database

The EROP database serves as a comprehensive national resource for oncological data, enabling robust epidemiological research through its standardized data collection forms, nationwide coverage, and multi-tiered quality control processes managed by governmental authorities. Its primary strength lies in the comparability of oncological indicators across regions, supported by uniform protocols and centralized oversight. The use of ICD-10 codes as the data entry basis further enhances the database's utility, allowing seamless alignment with international study designs and facilitating cross-border research collaborations.

Despite its strengths, EROP remains constrained by incomplete data capture in some regions, variability in reporting quality, and limited integration with non-oncology clinical systems. Data on behavioural risk factors, socioeconomic determinants, and patient-reported outcomes are currently lacking, restricting opportunities for comprehensive risk modelling. Inconsistencies in data synchronization between EROP forms also create gaps in socio-demographic information when merging records from different care settings. For instance, patients receiving hospital or outpatient treatment may not appear in dispensary surveillance records, leading to fragmented datasets. Additionally, therapy documentation is limited to hospital-administered treatments, excluding at-home care, and resulting in an incomplete representation of therapeutic practices. Addressing these gaps—through standardisation, capacity building, and digital integration—will be critical for realising the full potential of EROP as a research platform. Future directions may include linkage to mortality registries, natural language processing for unstructured clinical notes, and patient-facing platforms to capture quality of life and treatment experience.

The EROP database is also subject to biases common to cancer registries. Sampling bias arises because the UNEHS system, which feeds into the registry, covers only 36.4% of Kazakhstan's adult population [8], meaning the registry's cases may not fully represent the country's oncological patient population. This partial coverage risks excluding demographic groups with limited healthcare access, potentially skewing socio-demographic analyses. Ascertainment bias further complicates the data, as the registry predominantly captures advanced-stage cancers linked to hospitalizations and outpatient visits, underrepresenting early-stage or less symptomatic cases managed outside formal care settings. The registry's large sample size, while enhancing statistical power, can amplify both biases and the likelihood of detecting statistically significant relationships, potentially leading to spurious associations or overinterpretation of minor effects without rigorous methodological safeguards [15]. Therefore, rigorous methodological designs should be employed when using EROP in epidemiological studies

Conclusion

In summary, EROP serves not only as a clinical database for managing patient information but also as a valuable research tool in cancer epidemiology, enabling the development of evidence-based, population-specific health policies in Kazakhstan. It provides a

unique opportunity to investigate regional determinants of cancer and to analyze oncological burden across spatial and temporal dimensions. Ensuring the ethical use of registry data and conducting validation studies of EROP and other UNEHS databases are critical steps in strengthening this resource. Ongoing improvements in data quality, data entry procedures, and registry coverage, combined with methodological rigor, will further enhance the value and effectiveness of the national cancer registry in supporting public health research and guiding cancer control strategies in Kazakhstan.

Author Contributions: Conceptualization, A. B., A. A., and A. G.; methodology, A. B., A. A., R. A., D. A., and A. G.; validation, A. Y. and A. G.; formal analysis, A. A., A. B., and A. G.; writing – original draft preparation, A. A. and A. B.; writing – review and editing, A. B., D. A, and R. A.; supervision, A. G. All authors have read and agreed to the published version of the manuscript.
Disclosures: There is no conflict of interest for all authors.

Acknowledgments: None.

Funding: This research was funded by grants from the Ministry of Science and Higher Education of the Republic of Kazakhstan 2024–2026 (IRN: AP23484725; Title: Innovative Assessment of Solid Cancers Burden in Kazakhstan and Its 10-year Trend Using Big Healthcare Data and Forecasting Models). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval: The study was approved by the Institutional Review Board of Nazarbayev University (protocol code: NU-IREC 651/24112022 and date of approval: 28 November 2022), with exemption from informed consent.

References

1. World Health Organization. Noncommunicable diseases. 2024. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (accessed May 13, 2025).

2. Walsh TS, Stanworth S, Boyd J, Hope D, Hemmatapour S, Burrows H, Campbell H, Pizzo E, Swart N, Morris S. The age of blood evaluation (ABLE) randomised controlled trial: Description of the UK-funded arm of the international trial, the UK cost-utility analysis and secondary analyses exploring factors associated with health-related quality of life and health-care costs during the 12-month follow-up. Health Technol Assess (Rockv) 2017; 21:1 118. https://doi.org/10.3310/hta21620.

3. World Health Organization. Global tuberculosis control 2011:246. https://iris.who.int/handle/10665/44728 (accessed May 13, 2025).

4. Hotez PJ, Alibek K. Central Asia's hidden burden of neglected tropical diseases. PLoS Negl Trop Dis 2011;5. https://doi.org/10.1371/journal.pntd.0001224.

5. Aringazina A, Gulis G, Allegrante JP. Public Health Challenges and Priorities for Kazakhstan. Cent Asian J Glob Health 2012;1. https://doi.org/10.5195/cajgh.2012.30.

6. World Bank Group. The World Bank in Kazakhstan. Https://WwwWorldbankOrg/En/Country/Kazakhstan/Overview#1 n.d. https://www.worldbank.org/en/country/kazakhstan/overview#1.

7. Akhmedullin R, Aimyshev T, Zhakhina G, Yerdessov S, Beyembetova A, Ablayeva A, Biniyazova A, Seyil T, Abdukhakimova D, Segizbayeva A, Semenova Y, Gaipov A. In-depth analysis and trends of cancer mortality in Kazakhstan: a joinpoint analysis of nationwide healthcare data 2014-2022. BMC Cancer 2024; 24:1340. https://doi.org/10.1186/s12885-024-13128-2.

8. Gusmanov A, Zhakhina G, Yerdessov S, Sakko Y, Mussina K, Alimbayev A, Syssoyev D, Sarria-Santamera A, Gaipov A. Review of the research databases on population-based Registries of Unified electronic Healthcare system of

Kazakhstan (UNEHS): Possibilities and limitations for epidemiological research and Real-World Evidence. Int J Med Inform 2023;170. https://doi.org/10.1016/j.ijmedinf.2022.104950.

9. Midlenko A, Mussina K, Zhakhina G, Sakko Y, Rashidova G, Saktashev B, Adilbay D, Shatkovskaya O, Gaipov A. Prevalence, incidence, and mortality rates of breast cancer in Kazakhstan: data from the Unified National Electronic Health System, 2014–2019. Front Public Health 2023;11. https://doi.org/10.3389/fpubh.2023.1132742.

10. Mussina K, Kuanova B, Syssoyev D, Gaipov A, Poddighe D, Shaikhyzada K, Aimyshev T, Galiyeva D. Epidemiology of pediatric hematological malignancies in Kazakhstan: Data from Unified National Electronic Healthcare System 2014–2021. Eur J Pediatr 2024; 183:1683–91. https://doi.org/10.1007/s00431-023-05412-3.

11. Sakko Y, Aimagambetova G, Terzic M, Ukybassova T, Bapayeva G, Gusmanov A, Zhakhina G, Zhantuyakova A, Gaipov A. The Prevalence, Indications, Outcomes of the Most Common Major Gynecological Surgeries in Kazakhstan and Recommendations for Potential Improvements into Public Health and Clinical Practice: Analysis of the National Electronic Healthcare System (2014–2019). Int J Environ Res Public Health 2022;19. https://doi.org/10.3390/ijerph192214679.

12. Shatkovskaya O, Kaidarova D, Ongarbayev B, Sagi M, Tsimafeyeu I. Five-year overall survival of patients with advanced bladder cancer in Kazakhstan: OSURK registry study. vol. 11. 2023.

13. Umurzakov KT, Shalgumbayeva GM, Kaydarova DR, Smail Y, Ibrayev A, Sagidullin SO, Semenova YM. Epidemiological Characteristics of Male Reproductive Cancers in the Republic of Kazakhstan: Ten-Year Trends. vol. 51. 2022. https://doi.org/10.18502/ijph.v51i8.10265.

14. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: A cautionary note on the potential for bias. Clin Transl Sci 2014;7:342–6. https://doi.org/10.1111/cts.12178.

Epidemiology & Health Data Insights (eISSN 3080-8111)

Review Article. *Epidemiol Health Data Insights. 2025;1(1):ehdi005* https://doi.org/10.63946/ehdi/16411

Occupational Hazards in the Mining Industry: A Focus on Chrysotile Asbestos Exposure and Health Outcomes

Mohammad Abdullah Al Zubair Naim[1]* & Raju Ahmed[2]*

[1]*Division of Nephrology, Department of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States*

[2]*College of Public Health, University of Iowa, Iowa City, IA, United States*

** These authors have equally contributed to this article.*

Received: May 20, 2025
Accepted: May 27, 2025

Corresponding author:

Mohammad Abdullah Al Zubair Naim

Email mnaim@uthsc.edu

This work is licensed under a **Creative Commons Attribution 4.0 International License**

Abstract: Notwithstanding notable advancements in occupational health and safety over the past century, occupational injuries and illnesses occur frequently worldwide. Mining is among the highest-risk occupational environments, exposing workers to acute and chronic health problems. This narrative review explores research evidence of different mining-related hazards, with a special focus on asbestos exposure. Mineworkers more commonly face biological, chemical, physical, ergonomics, and psychological hazards than other industries. Exposure to respiratory particles like silica and coal dust contributes to respiratory diseases like pneumoconiosis, silicosis, chronic obstructive pulmonary disease (COPD), and other lung diseases. Due to the physically demanding nature of the work and poor ergonomics, the risk of musculoskeletal disorders (MSDs) occurring in soft tissues and surrounding body structures is high in mineworkers. Despite technological advancement and safety precautions, the mining industry remains a high-risk workplace with disproportionately higher rates of occupational injuries. Routine exposure to carcinogens results in a high risk of stomach and lung cancers. Chrysotile is the most used asbestos in the mining industry, causing asbestosis and mesothelioma. While some countries have introduced regulations to limit or phase out the use of asbestos, asbestos is still used in many countries. Studies are warranted to investigate the immediate and long-term health effects and potential risk factors. Essential preventive measures, consistent enforcement of safety standards, and accessible healthcare are essential to ensure overall well-being.

Keywords: Occupational health; mineworkers; asbestos exposure; chrysotile; respiratory diseases; occupational cancer; workplace safety; mining injuries

Introduction

Occupational illnesses include physical or psychological illnesses that are primarily caused by the work environment or related activities, often resulting in significant disruption, disability, or death (1). A wide range of health problems can be caused or worsened by work conditions, which include respiratory issues like asthma and pneumoconiosis, musculoskeletal disorders, mental health challenges such as stress and depression, and various forms of cancers (2). Despite notable advancements in occupational health and safety over the past century, an estimated 317 million nonfatal occupational injuries and 321,000 occupational fatalities still occur globally each year (3). These injuries and illnesses vary depending on the type of occupation and the specific hazards involved, such as chemical, biological, or physical hazards (4, 5).

Mining is one of the highest-risk occupational environments characterized by numerous hazards, exposing workers to conditions that can lead to both acute and chronic health problems (5). Generally, mineworkers have a significantly higher prevalence of occupational diseases compared to the general working population, including exposure to harmful particulates such as silica dust, coal dust, metals-containing dust, diesel particulate matter, and industrial noise (6). The working conditions in mines, including exposure to dust, chemicals, physical strain, and poor ergonomic practices, contribute significantly to the higher morbidity and mortality rates in mining workers (5). Mining operations also expose workers to risks such as rock falls, equipment accidents, and exposure to harmful substances. In addition to that, the continued use of outdated equipment and mining techniques exacerbates the risk of accidents and toxic exposures (7). The risks associated with mining can vary considerably depending on several factors, including the type of mining (e.g., underground vs. surface), the mineral being extracted (e.g., coal, copper), exposure levels and their temporal change, and the geographical context in which the operation occurs (6).

Despite the regulatory frameworks and safety protocols in the mining sector, incidents remain common and occur at significantly higher rates compared to other employment sectors (8). Key determinants that influence safety in mining include geological complexity, regulatory enforcement, technological development, and environmental conditions such as extreme temperatures or seismic activity. Studies found that mines with robust safety regulations and modern technologies generally experience lower injury rates (9). Therefore, understanding and addressing the specific risks associated with different mining environments is crucial for improving the safety of mineworkers.

This review aims to explore the evidence obtained from research on mining-related injuries, with a special focus on asbestos exposure. The review highlights the need for improved safety measures, worker training, and enforcement of regulations to mitigate risks and enhance workplace safety in the mining industry.

Occupational hazards in mining industry

Occupational hazards fall into several categories-biological, chemical, physical, ergonomics, and psychological (5), with a variable magnitude of prevalence across different sectors, regions, and individuals. While every occupation entails the risk of injury from occupational hazards, the mining industry is one of the most hazardous occupations compared to other sectors (6).

Although the mining industry accounts for just 1% of the global workforce, it accounts for approximately 8% of all fatal workplace accidents (10). Mineworkers experience some of the highest rates of occupational injuries and illnesses across all industries, primarily due to their inherently hazardous and labor-intensive physical environments and exposure to toxic substances. Compared to the general working population, the mining industry has a higher percentage of manual laborers, putting them at different occupational exposures than occupations. A national study by Robinson et al. (2023), which compared mineworkers to six other groups of manual laborers-based industries in the United States, found that mineworkers had a higher crude prevalence of hypertension, moderate-to-severe hearing loss, lower back, neck, leg, and joint pain (6).

Common mining-related health issues and injuries

Respiratory diseases

Exposures to various respirable particulates have been linked to chronic occupational diseases in mineworkers, with excessive inhalation of silica and coal dust, as well as other particulates, contributing to various respiratory diseases and health-related issues. A study of more than 2,500 former US coal mineworkers found that over half had abnormal lung function, often due to prolonged exposure to coal dust (11).

Pneumoconiosis, commonly called 'black lung disease,' is a dust-induced occupational lung disease prevalent among mineworkers, particularly those working in coal mines (12). A 23-year follow-up study on US coal mineworkers found a direct correlation between the duration of coal dust exposure and the severity of pneumoconiosis, confirming that more prolonged exposure significantly increased disease progression and mortality risk (13).

Mineworkers exposed to high levels of silica dust are at a significantly increased risk of developing silicosis, a progressive disease that often leads to debilitating respiratory failure (14). A cross-sectional study by Cowie et al. (1991) investigating gold mineworkers in South Africa reported a higher prevalence of silicosis-associated pulmonary dysfunction, with dyspnea on exertion linked to the duration of underground exposure and the concentration of dust exposure in the workplace (15).

Another respiratory disorder with a notably higher prevalence among mineworkers than the general population is chronic obstructive pulmonary disease (COPD), which is developed as a result of long-term exposure to mining dust and fumes. Other reported respiratory symptoms and disorders include bronchitis, chronic bronchitis, chronic cough, and chronic phlegm (16). Table 1 furnishes the frequencies of different mine dust lung diseases reported by Resources Safety and Health Queensland (Australia) in 2024.

Mineworkers in the US, particularly coal mineworkers, were found to have increased odds of death from pneumoconiosis, lung cancer, and COPD compared to the rest of the population, caused by the respirable coal mine dust (17).

Musculoskeletal Disorders

Musculoskeletal disorders (MSDs) are soft tissue disorders and surrounding structures in the neck, shoulders, elbows, wrists, and lower back (18). These are among the most widespread occupational health problems in the mining industry (19). MSDs accounted for 32% of all non-fatal injuries among full-time US workers in 2014 (20). A meta-analysis by Rabiei et al. (2021) examining the prevalence of MSDs among mineworkers around the world reported MSDs to be one of the most common reported occupation-related disorders, with a prevalence of 50.39% in the upper back and 16.03% in the knees (19). Over 65% of underground coal mineworkers in Eastern India complained of MSD pain, with the highest, approximately 58%, of the pain identified in the lower back (21).

There are a number of occupational risk factors contributing to the development of MSDs in mining workers due to the physically demanding nature of mining work, such as poor ergonomics, including repetitive movements, heavy lifting, and awkward postures (22). The use of vibrating tools and machinery in mining is another significant risk factor for MSDs. Vibration and ergonomic exposures were found to be associated with MSDs of the neck and shoulder (22). It was also reported that prolonged exposure to hand-arm vibration is associated with conditions such as hand-arm vibration syndrome and rotator cuff syndrome among mineworkers (22).

Occupational Injuries

The mining industry, despite advancements in workplace safety, continues to remain a high-risk workplace with disproportionately higher rates and severity of occupational injuries (7). The physical environment of mines is characterized by confined spaces, heavy machinery, and the potential for structural failures, contributing to the higher likelihood of occupational injuries.

The US Bureau of Labor Statistics (2023) reported incidence rates for non-fatal occupational injuries in the mining sectors to be 130 per 10,000 full-time employees in 2023 (23). An investigation into the Australian mining sector by Jones et al. (2013) in 2019 found that there were a total of 837 reported injuries, including 658 severe cases (three fatalities) and 179 minor injuries, highlighting the ongoing risks mineworkers face (24). The authors reported sprains and strains (69%) as the most recorded injury, followed by fractures (10%), lacerations (6%), crushing injuries (5%), bruises and contusions 4%, and dislocations and displacements 2%. Other injuries include amputations, punctures, chemical effects, bites, foreign bodies, loss of consciousness, thermal burns, and flash and arc burns (24).

A decade-long investigation by Chen et al. (2012) that examined the trends of coal mine accidents in China, although reported a significant decrease in accidents and fatalities after 2005 compared to earlier periods, the absolute number of registered coal mine fatalities in 2010 was still high (n= 2,433) (25).

Notwithstanding technological advancements and stricter safety protocols, mining remains one of the most injury-prone industries globally, highlighting the need for improved hazard control, comprehensive worker training, and effective emergency preparedness.

Cancers

Mineworkers are routinely exposed to a wide range of hazardous substances, many of which are known or suspected carcinogens. Long-term exposure to these agents significantly elevates the risk of various cancers. A meta-analysis by Alif et al. (2022), investigating cancer and mortality in coal mine workers, reported

an increased risk of stomach cancers among mineworkers (standardized mortality ratio: 1.11, 95% CI 0.97 to 1.35) (26).

In addition, mineworkers are at a higher risk of developing lung cancers compared to the general population. A large Scandinavian registry study identified mineworkers as having a high occupational risk for lung cancer, largely due to exposure to silica dust and radon daughters, as well as other airborne carcinogens such as arsenic, chromium (VI), asbestos, polycyclic aromatic hydrocarbons (PAHs), diesel exhaust, and nickel (27).

However, epidemiological studies reported conflicting results on lung cancer among coal mineworkers. Hosgood and colleagues investigated the association between coal mining and lung cancer in China, observing an elevated lung cancer risk among coal mineworkers compared to non-coal miners (28). On the contrary, Brown and colleagues, who investigated the occurrence of cancer in a cohort of coal mineworkers in Australia, suggested that the coal industry in the New South Wales does not appear to be a general risk of cancer (29). Such discrepancies underscore the importance of modifiable factors. The type of mineral extracted, level and duration of exposure, and co-factors such as smoking play a critical role in modulating cancer risk in mining populations. Smoking, in particular, is known to have a synergistic effect with workplace carcinogens, amplifying the overall risk of respiratory cancers (30). Studies demonstrated smoking mediated the relationship between asbestos and lung cancer among mineworkers (31).

These findings highlight the need for region-specific occupational health monitoring and targeted cancer prevention strategies for mineworkers.

Chrysotile asbestos exposure and consequences

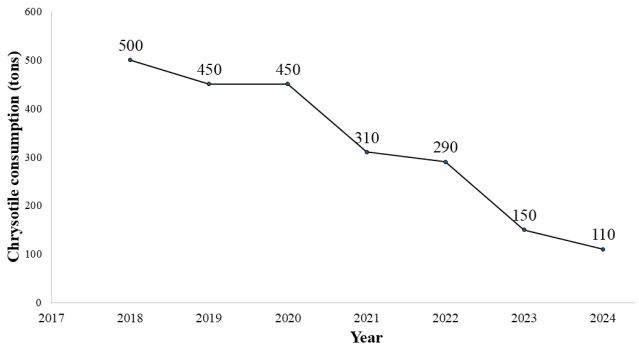
One of the major occupational hazards in the mining industry is airborne environmental pollution within the mines, primarily in the form of toxic dust, gases, fumes, and vapors. Among these, the inhalation of fine particulate matter, especially asbestos fibers, poses a significant health risk to miners. Asbestos refers to a group of naturally occurring fibrous minerals that have been widely used in industrial applications for their durability, heat resistance, and insulating properties. Chrysotile, known as white asbestos, is the most commercially used asbestos globally (32).

The mining and processing of asbestos have had severe health consequences for mineworkers, particularly those who are frequently exposed to airborne asbestos fibers during extraction and processing. Fibers deposited in the lung parenchyma lead to a progressive inflammation of the tissues, resulting in a fibrotic response that impairs gas exchange, resulting in asbestosis, a potentially fatal lung disorder that may cause progressive dyspnea and respiratory failure (33).

Chrysotile belongs to the serpentine mineral group and is classified as a carcinogen by international health authorities due to its strong association with mesothelioma, a rare and aggressive cancer affecting the mesothelium, the protective lining of several internal organs (34). Mesothelioma is often diagnosed at an advanced stage, which limits the treatment options and results in poor survival. A study on cancer mortality of a historical cohort of 30,445 mineworkers who worked for more than 30 years in the chrysotile mine and its enrichment factories in Russia confirmed that exposure to dust containing chrysotile increased the risk of cancer development in a dose-dependent manner (35).

Despite the well-established relationship between asbestos and mesothelioma, chrysotile is found in 95% of all asbestos mined today (32). However, global chrysotile consumption has significantly declined due to increased awareness of its health risks and regulatory bans in many countries. For example, figure 1 shows the domestic chrysotile consumption trend in the United States between 2018 to 2024 (36).

Figure 1: Domestic chrysotile consumption trend in United States between 2020 to 2024



Global policy initiatives in reducing asbestos-related health risks

Several countries have implemented successful policies to mitigate the occupational health risks associated with asbestos, particularly in the mining and construction sectors. These examples can inform more effective regulations in regions where asbestos exposure remains a significant concern:

1. *Australia:* A complete ban on all forms of asbestos was implemented in 2003. Prior to the ban, the country had already phased out the use of asbestos in construction materials and introduced national standards for asbestos removal. Post-ban, strict licensing systems for asbestos handling and removal, along with public awareness campaigns, significantly reduced occupational exposure and related diseases (37).
2. *European Union (EU):* In 2005, the EU banned chrysotile asbestos after years of concern over its health risks. Before that, member states such as France, Germany, and Sweden had already enacted national bans and initiated asbestos abatement programs in schools, public buildings, and older industrial sites. Additionally, the EU established directives for asbestos waste management to ensure the safe disposal of contaminated materials, thereby reducing the risks of secondary exposure to asbestos fibers (38, 39).
3. *Japan:* In the 1990s, Japan adopted a phased approach to banning asbestos, culminating in a total ban in 2012. In response to growing public concern and the surge in mesothelioma cases, they also introduced a compensation system in 2006 to support both occupationally and environmentally exposed individuals. Additionally, the government made significant investments in retrofitting public infrastructure to ensure the safe removal of asbestos-containing materials, aiming to prevent future exposures and protect public health (40).
4. *South Korea:* South Korea implemented a complete ban on asbestos in 2009. The national asbestos surveillance system in South Korea includes a registry for asbestos-exposed individuals and mandatory health monitoring for workers with a history of employment in asbestos-related industries. These initiatives have significantly strengthened early disease detection and long-term monitoring, enhancing prevention and compensation efforts for affected populations (41).
5. *Brazil:* In 2017, Brazil instituted a nationwide ban on commercial asbestos use. Authorities also began encouraging industries to adopt safer alternative materials and provided retraining programs to support workers transitioning out of asbestos-related jobs. These steps aimed to reduce future exposure risks while supporting affected communities through economic and occupational adjustments (42).

Occupational hazards: beyond physical harm

In the mining industry, some health problems develop silently over time upon long-term exposure to toxic dust and chemicals, severely impacting the daily well-being of workers. Diseases such as asbestosis and lung cancer lead to constant breathlessness, fatigue, and pain, making everyday tasks a challenge. These physical struggles often coexist with mental health issues like anxiety and depression (43).

Living with a long-term illness becomes financially overwhelming. Medical treatments can be expensive, and when workers are too sick to continue working, the loss of income puts enormous economic pressure on them and their families. Over time, many affected individuals become socially isolated, impairing their emotional health and quality of life (44).

Mineworkers facing long-term illnesses require support like access to counseling, mental health care, and programs that help them return to social life and work. Community-based efforts such as peer support groups, workplace mental health screenings, and financial assistance can make a real difference in helping affected workers cope and recover (45, 46).

Table 1. Reports of mine dust lung diseases (all mining) reported in 2024 in Queensland, Australia (47)

Lung disease	Cases (n)
Chronic Obstructive Pulmonary Disease (COPD)	101
Pneumoconiosis	8
Silicosis	15
Cancer	3
Other lung diseases	9
Multiple lung diseases	27

Conclusion

While the mining industry contributes significantly to global economies, it continues to pose substantial health and safety challenges for its workforce. Despite the improvement in the safety of the mining environment, the nature of the mining industry exposes mineworkers to various physical and environmental risks, potentially compromising their quality of life. Prospective studies are warranted to investigate the immediate and long-term physical and mental health effects and potential risk factors associated with those. Studies should also explore the complex interplay of physical, chemical, and organizational risks of mining workers. While some countries have introduced regulations to limit or phase out asbestos use, many permit it in various industries, including mining. Therefore, studies should underscore asbestos-related health risks in mining populations to influence policy decisions. Strategies to better protect the mineworkers include modernizing mining equipment, providing adequate personal protective equipment, and implementing regular medical screenings to detect health issues at an early, more treatable stage. Essential preventive measures, consistent enforcement of safety standards, and accessible healthcare will likely protect mineworkers from the potential and detrimental effects of occupational illnesses and improve their overall well-being.

Author Contributions: Conceptualization, M.N., R.A.; methodology, M.N., R.A.; writing – original draft preparation, M.N., R.A.; writing – review and editing, M.N., R.A. All authors have read and agreed to the published version of the manuscript.

Disclosures: There is no conflict of interest for all authors.

Acknowledgments: None.

Funding: None.

Ethics approval: Not applicable.

References

1. Leigh JP, Markowitz SB, Fahs M, Shin C, Landrigan PJ. Occupational injury and illness in the United States. Estimates of costs, morbidity, and mortality. Arch Intern Med. 1997;157(14):1557-68.

2. Peters S, Undem K, Solovieva S, Selander J, Schlünssen V, Oude Hengel KM, et al. Narrative review of occupational exposures and noncommunicable diseases. Ann Work Expo Health. 2024;68(6):562-80.

3. Guercia G. ILO calls for urgent global action to fight occupational diseases, Geneva, 2013.

4. Knecht H, Balanay JAG, Langley R, Tutor Marcom R, Richards SL. Systematic Review of Biological, Chemical, Ergonomic, Physical, and Psychosocial Hazards Impacting Occupational Health of United States Forestry Workers. Journal of Forestry. 2023;122(2):159-70.

5. Donoghue AM. Occupational health hazards in mining: an overview. Occup Med (Lond). 2004;54(5):283-9.

6. Robinson T, Sussell A, Scott K, Poplin G. Health conditions among male workers in mining and other industries reliant on manual labor occupations: National Health Interview Survey, 2007–2018. Am J Ind Med. 2023;66(8):692-704.

7. Kecojevic V, Komljenovic D, Groves W, Radomsky M. An analysis of equipment-related fatal accidents in US mining operations: 1995–2005. Saf Sci. 2007;45(8):864-74.

8. Cruz-Ausejo L, Cama-Ttito NA, Solano PF, Copez-Lonzoy A, Vera-Ponce VJ. Occupational accidents in mining workers: scoping review of studies published in the last 13 years. BMJ Open. 2024;14(10):e080572.

9. Onifade M, Said KO, Shivute AP. Safe mining operations through technological advancement. Process Safety and Environmental Protection. 2023;175:251-8.

10. Baraza X, Cugueró-Escofet N, Rodríguez-Elizalde R. Statistical analysis of the severity of occupational accidents in the mining sector. Journal of Safety Research. 2023;86:364-75.

11. Go LH, Almberg KS, Rose CS, Zell-Baran LM, Harris DA, Tomann M, et al. Prevalence and severity of abnormal lung function among US former coal miners with and without radiographic coal workers' pneumoconiosis. Occupational and environmental medicine. 2022;79(8):527-32.

12. Kirsten Almberg RC. Modern Coal Miners Have Higher Death Rates From Lung Diseases Than Their Predecessors blogs.cdc.gov: Centers for disease control and prevention; 2025 [updated November 25, 2024. Available from: https://blogs.cdc.gov/niosh-science-blog/2023/02/27/mining-lung-disease/.

13. Attfield M, Kuempel E. Mortality among US underground coal miners: A 23 ‐ year follow ‐ up. Am J Ind Med. 2008;51(4):231-45.

14. Steenland K, Ward E. Silica: a lung carcinogen. CA Cancer J Clin. 2014;64(1):63-9.

15. Cowie RL, Mabena SK. Silicosis, chronic airflow limitation, and chronic bronchitis in South African gold miners. Am Rev Respir Dis. 1991;143(1):80-4.

16. Ross MH, Murray J. Occupational respiratory disease in mining. Occupational Medicine. 2004;54(5):304-10.

17. Almberg KS, Halldin CN, Friedman LS, Go LH, Rose CS, Hall NB, et al. Increased odds of mortality from non-malignant respiratory disease and lung cancer are highest among US coal miners born after 1939. Occupational and environmental medicine. 2023;80(3):121-8.

18. Punnett L, Wegman DH. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. J Electromyogr Kinesiol. 2004;14(1):13-23.

19. Rabiei H, Malakoutikhah M, Vaziri MH, Sahlabadi AS. The prevalence of musculoskeletal disorders among miners around the world: a systematic review and meta-analysis. Iran J Public Health. 2021;50(4):676.

20. BLS U. Nonfatal occupational injuries and illnesses requiring days away from work, 2011. 11/08/2012; 2012.

21. Bandyopadhyay A, Dev S, Gangopadhyay S. A study on the prevalence of musculoskeletal disorders among the coalminers of Eastern Coalfields of India. International Journal of Occupational Safety and Health. 2012;2(2):34-7.

22. Charles LE, Ma CC, Burchfiel CM, Dong RG. Vibration and Ergonomic Exposures Associated With Musculoskeletal Disorders of the Shoulder and Neck. Safety and Health at Work. 2018;9(2):125-32.

23. Statistics BoL. EMPLOYER-REPORTED WORKPLACE INJURIES AND ILLNESSES – 2023 www.bls.gov/iif 2024.

24. Jones R, Cattani M, Cross M, Boylan J, Holmes A, Boothroyd C, et al. Serious injuries in the mining industry: preparing the emergency response. Australasian Journal of Paramedicine. 2019;16:1-6.

25. Chen H, Qi H, Long R, Zhang M. Research on 10-year tendency of China coal mine accidents and the characteristics of human factors. Saf Sci. 2012;50(4):745-50.

26. Alif SM, Sim MR, Ho C, Glass DC. Cancer and mortality in coal mine workers: a systematic review and meta-analysis. Occup Environ Med. 2022;79(5):347-57.

27. Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparen P, Tryggvadottir L, et al. Occupation and cancer - follow-up of 15 million people in five Nordic countries. Acta Oncol. 2009;48(5):646-790.

28. Hosgood III HD, Chapman RS, Wei H, He X, Tian L, Liu LZ, et al. Coal mining is associated with lung cancer risk in Xuanwei, China. Am J Ind Med. 2012;55(1):5-10.

29. Brown AM, Christie D, Taylor R, Seccombe MA, Coates MS. The occurrence of cancer in a cohort of New South Wales coal miners. Aust N Z J Public Health. 1997;21(1):29-32.

30. Taeger D, Pesch B, Kendzia B, Behrens T, Jockel KH, Dahmann D, et al. Lung cancer among coal miners, ore miners and quarrymen: smoking-adjusted risk estimates from the synergy pooled analysis of case-control studies. Scand J Work Environ Health. 2015;41(5):467-77.

31. Ngamwong Y, Tangamornsuksan W, Lohitnavy O, Chaiyakunapruk N, Scholfield CN, Reisfeld B, et al. Additive Synergism between Asbestos and Smoking in Lung Cancer Risk: A Systematic Review and Meta-Analysis. PLoS One. 2015;10(8):e0135798.

32. Linton A, Vardy J, Clarke S, van Zandwijk N. The ticking time-bomb of asbestos: its insidious role in the development of malignant mesothelioma. Crit Rev Oncol Hematol. 2012;84(2):200-12.

33. Musk AW, de Klerk N, Reid A, Hui J, Franklin P, Brims F. Asbestos-related diseases. Int J Tuberc Lung Dis. 2020;24(6):562-7.

34. Arsenic, metals, fibres, and dusts. IARC Monogr Eval Carcinog Risks Hum. 2012;100(Pt C):11-465.

35. Schuz J, Kovalevskiy E, Olsson A, Moissonnier M, Ostroumova E, Ferro G, et al. Cancer mortality in chrysotile miners and millers, Russian Federation: main results (Asbest Chrysotile Cohort-Study). J Natl Cancer Inst. 2024;116(6):866-75.

36. Center NMI. Asbestos Statistics and Information: U.S. Geological Survey; 2025 [Available from: https://www.usgs.gov/centers/national-minerals-information-center/asbestos-statistics-and-information.

37. Soeberg M, Vallance DA, Keena V, Takahashi K, Leigh J. Australia's Ongoing Legacy of Asbestos: Significant Challenges Remain Even after the Complete Banning of Asbestos Almost Fifteen Years Ago. Int J Environ Res Public Health. 2018;15(2).

38. Valić F. The asbestos dilemma: II. The ban. Arh Hig Rada Toksikol. 2002;53(3):203-11.

39. Westerholm P, Remaéus B, Svartengren M. The tale of Asbestos in Sweden 1972–1986—the pathway to a Near-Total Ban. International Journal of Environmental Research and Public Health. 2017;14(11):1433.

40. Furuya S, Takahashi K. Experience of Japan in Achieving a Total Ban on Asbestos. Int J Environ Res Public Health. 2017;14(10).

41. Jung SH, Kim HR, Koh SB, Yong SJ, Chung MJ, Lee CH, et al. A decade of malignant mesothelioma surveillance in Korea. Am J Ind Med. 2012;55(10):869-75.

42. Algranti E, Santana VS, Campos F, Salvi L, Saito CA, Cavalcante F, et al. Analysis of mortality from asbestos-related diseases in Brazil using multiple Health Information systems, 1996–2017. Safety and Health at Work. 2022;13(3):302-7.

43. Zawada K, Bratek A, Krysta K. Psychological distress and social factors in patients with asthma and chronic obstructive lung disease. Psychiatria Danubina. 2015;27(suppl 1):462-4.

44. Padmanabhan M, Kadhiravan T, Rajaram M, Batmanabane G. A panoramic view of patients' beliefs and knowledge about chronic respiratory disease. Cureus. 2022;14(11).

45. Höhl W, Moll S, Pfeiffer A. Occupational therapy interventions in the treatment of people with severe mental illness. Current opinion in psychiatry. 2017;30(4):300-5.

46. Cano Prieto I, Simó Algado S, Prat Vigué G. Peer interventions in severe mental illnesses: a systematic review and its relation to occupational therapy. Occupational Therapy in Mental Health. 2023;39(2):99-136.

47. Queensland So. Mine dust lung diseases qld.gov.au: Queensland Government; 2025 [updated April 30, 2025. Available from: https://www.business.qld.gov.au/industries/mining-energy-water/resources/safety-health/mining/accidents-incidents-reports/mine-dust-lung-diseases.

Epidemiology & Health Data Insights (eISSN 3080-8111)

Methodological Notes. *Epidemiol Health Data Insights. 2025;1(1):ehdi001* https://doi.org/10.63946/ehdi/16216

Overcoming the Misinterpretation of a Measure in Observational Studies: What and When?

Ruslan Akhmedullin [1]

[1]School of Medicine, Nazarbayev University, Astana, Kazakhstan

Received: March 27, 2025
Accepted: April 28, 2025

Corresponding author:
ruslan.akhmedullin@nu.edu.kz

This work is licensed under a **Creative Commons Attribution 4.0 International License**

Abstract: The measure of association in observational studies is often prone to misinterpretation, further necessitating a discussion of methodological challenges. Understanding the underlying causes and proposing proper steps may help prevent such issues in the literature. This article provides straightforward explanations of terminology and a decision tree to select appropriate measures and accurate interpretations.

Keywords: Measure of Effect; Bias; Statistical Analysis;

Introduction

Research on health-related outcomes has always been of particular interest. Access to real-world data (RWD) provides new avenues for medical research that are fuelling evidence generation (1). However, a considerable amount of this evidence comes from observational studies, and in addition to the lack of internal validity, their estimates are often associated with misinterpretation (2, 3). As a result, their impact remains limited as long as decision makers face issues in their reporting practice.

The odds ratio (OR) is arguably one of the most commonly used measures of association and is also frequently interpreted incorrectly, pretending to represent the risk of outcome (4). It can, however, approximate the risk ratio (RR), but only if certain conditions are met. Furthermore, the incidence rate ratio (IRR) or hazard ratio (HR) can and should also be calculated for specific case-control designs, although the OR is often reported instead. It is worth noting that there is nothing wrong with the OR, as long as it is interpreted appropriately. Previous surveys (5, 6) of over 200 observational studies published in leading medical journals revealed systematic issues in the methodology and interpretation of their findings, most of which inaccurately reported their measures, lacking sufficient discussion of their assumptions. Unfortunately, two decades after these findings were published, one can see that things are not going any better (7). Hence, discussing underlying issues and proposing appropriate steps might become key to preventing the recurrence of such issue.

Short-walk through ideas behind case-control and cohort studies.

A case-control study represents a family of different designs (hybrid retrospective, case-base, case-exposure, matched, nested, case-crossover, and case-cohort) to measure the distribution of exposure among the study population (8). Until certain assumptions are met, it is unclear whether the main parameter is the odds of exposure or the disease itself. Some of these designs are particularly efficient when the prevalence of the outcomes is low. Such circumstances allow for different estimates (e.g., RR and OR) to be used interchangeably. However, ignoring this assumption leads to an exaggeration of the effect size, further introducing both misinterpretation and magnitude error. As illustrated in Figure 1, when the outcome is less prevalent, the line becomes more diagonal, making RR and OR somewhat similar. However, as the outcome became more common, the difference between the two measures increases.

The RWD granted access to longitudinal data and provided detailed information on both the exposure and outcome. Unlike case-control studies, cohort studies (prospective, retrospective) can explicitly relate exposure to an outcome with an intuitive interpretation. The widespread use of logistic regression has made the OR a common measure, even for longitudinal data. However, it does not have a simple interpretation of RR. Furthermore, common outcomes may require incorporation of modified regression models. Nevertheless, this assumption only relates to studies that sample at the end of the follow-up period from a specific cohort, and current evidence suggests that this situation is rarely applicable in practice (5).

To properly define whether the main parameter of interest is OR, RR, IRR, or HR, one should understand the insights from the study design, sampling strategy, source of population, statistical analysis, and their interplay. This must be where the most confusion comes from, leading to some ignorance of these factors, which in turn causes errors in analysis and interpretation. Here, we aimed to simplify the terminologies, relate them to each other, and develop a straightforward decision tree for the selection of the appropriate measure with a focus on statistical techniques. To maintain simplicity, we set aside discussions of censoring, competing risks, and loss of follow-up.

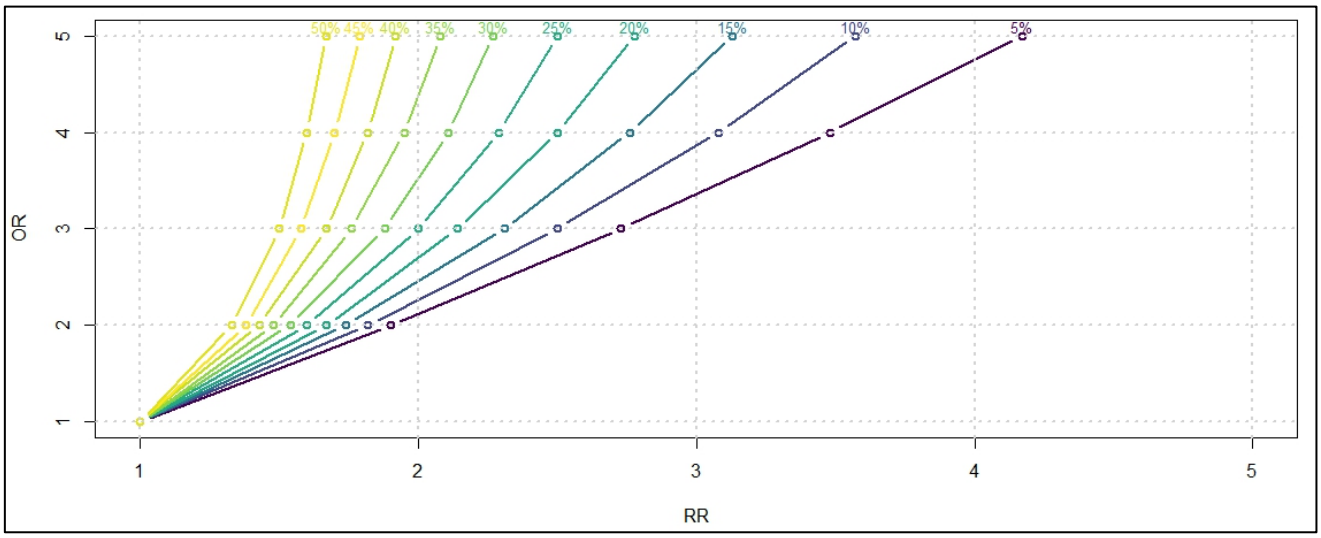


Figure 1. Comparison of odds ratios and risk ratios for different prevalence of outcome values.

The RR was approximated using formula: RR = OR / [(1−P0) + (P0 x OR). Where P0 represents the prevalence of the outcome in the non-exposed group.

Clearing confusion around the terminologies to define a measure of association

There are two population sources are available for observational studies: static and dynamic, also known as closed and open cohorts, respectively. The main difference lies in their structure over time. That is, in the closed cohort, participants are selected once, and no new entries are considered. For instance, a cohort of individuals diagnosed with pancreatic cancer in a particular year was further followed for over a year. However, open cohorts accept changes in size, follow-up, and multiple enrolments for the same participant. For example, a cohort of smokers can join if they are current smokers and leave if they quit smoking.

Both case-control and cohort studies require proper selection of the comparison group, and further interpretation of the OR differs depending on the nature of the denominator (person-time at risk, persons at risk, or survivors). The closed cohort has three commonly used sampling approaches. First, exclusive or cumulative incidence sampling that samples controls from those who remained free of outcome at the end of follow-up, where the denominator represents survivors. The measure for this sampling is OR of exposure. It can be interpreted as the RR of an outcome, assuming that the outcome is rare (<10%). Second, controls

selected from the total study population at the beginning of follow-up: inclusive sampling (case cohort). Its denominator represents the population at risk, and the measure of interest is also OR but should be interpreted as RR directly, regardless of the rare disease assumption (9). Third, controls can be sampled concurrently (longitudinally) with cases containing person-time at risk (sum of any time unit) in the denominator. The OR, in this instance will estimate the rate ratio, but it is more intuitive to use the IRR/HR right away for a clear interpretation of the estimand.

Within the open cohorts, controls are selected from the time unit at risk using incidence density sampling (nested case-control study). Another approach is to select controls at some point in time, either at the end, beginning, or during the period in which the cases are diagnosed. Such sampling allows the OR to better approximate the RR, compared to previous schemes, even assuming a rare disease (10). Although this instance allows the interpretation of OR as the rate ratio, it is more intuitive to use IRR/HR, given the information on the time unit for exposure (e.g., treatment onset, disease diagnosis date) and outcome. Finally, when it comes to recurrent diseases (e.g., acute conditions), cases return to the at-risk population after recovery (transient exposure) and represent their own controls (case-crossover). It allows the calculation of OR from conditional logistic regression, which ensures a proper account for such a dependency. The OR again approximates the RR if the outcome is rare and is not directly interchangeable.

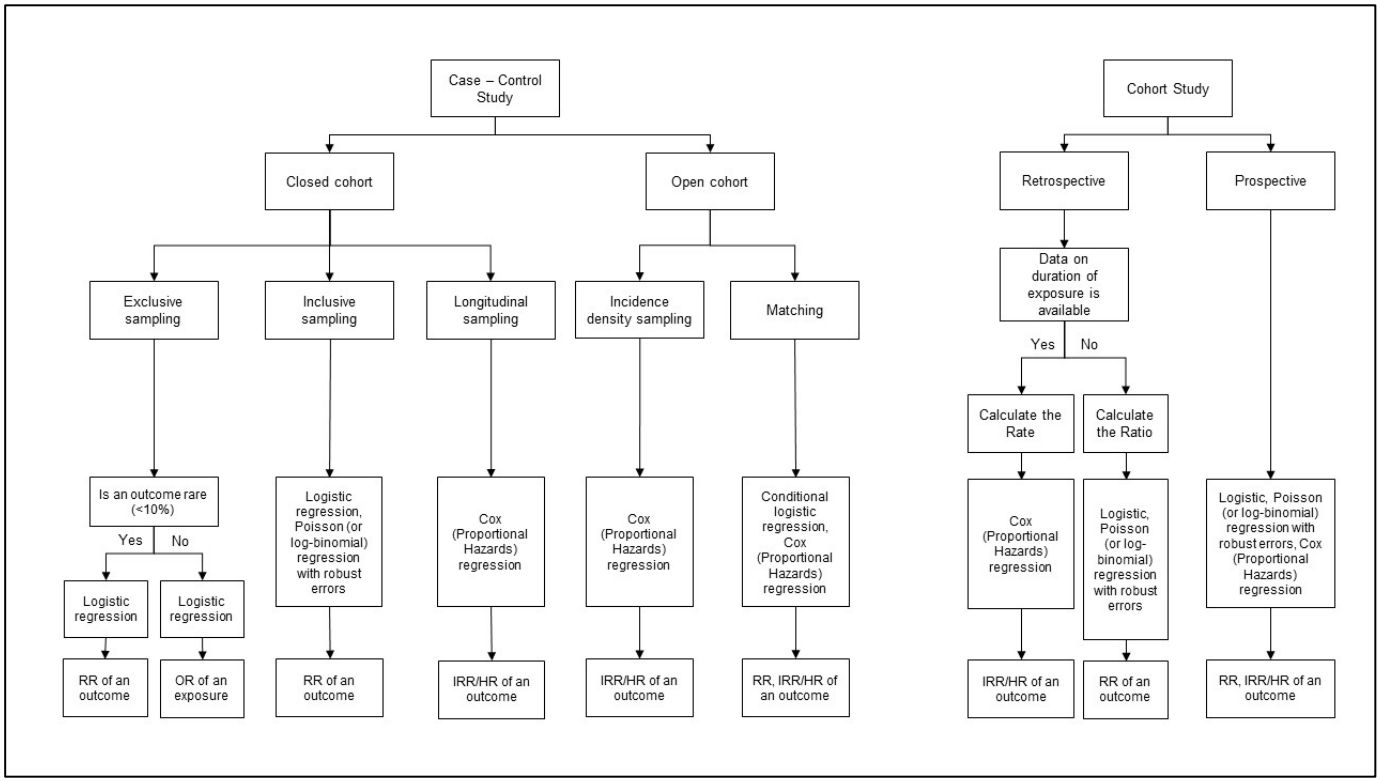


Figure 2. Decision tree for measure selection and interpretation.

In Figure 2, one can recognize that the only situation where the measure of association should be the OR, and interpreted as such, refers to a case-control study conducted within a closed cohort with exclusive sampling. Despite the fact that disease OR and exposure OR are identical, they should be interpreted as exposure OR, not a disease OR. When the study had a case-cohort design, the OR could not be interpreted as such but only as RR. The same is true for designs that use person-time rather than participant or survivor in the denominator. In these designs, the OR is not equal to the disease OR, although most studies that estimated the IRR/HR reported the OR instead. It is common to turn OR into RR, IRR, and HR immediately. However, the OR is questionable unless the assumptions are clarified.

Conclusion

Current literature suggests that a case-control study is nothing but a cohort study achieved by sampling a subset of controls to obtain a measure of exposure distribution among them. The discordance between the statistical interpretations of measures is likely driven by challenges in understanding methodological details in both epidemiology and statistics. To properly select and accurately interpret a measure, one should consider population, sampling, and statistical tests to avoid misinterpretation of the estimand.

References

1. Wang SV, Schneeweiss S, Franklin JM, Desai RJ, Feldman W, Garry EM, et al. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. JAMA. 2023;329(16):1376–85. doi:10.1001/jama.2023.4177

2. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. N Engl J Med. 1999;341(4):279–83. doi:10.1056/NEJM199907223410406

3. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ. 1998;316(7136):989–91. doi:10.1136/bmj.316.7136.989

4. Holcomb WL Jr, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and misuse of the odds ratio. Obstet Gynecol. 2001;98(4):685–8. doi:10.1016/S0029-7844(01)01488-0

5. Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol. 2008;168(9):1073–81. doi:10.1093/aje/kwn191

6. Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ. 2004;329(7471):883. doi:10.1136/bmj.38250.571088.55

7. Labrecque JA, Hunink MMG, Ikram MA, Ikram MK. Do Case-Control Studies Always Estimate Odds Ratios? Am J Epidemiol. 2021;190(2):318–21. doi:10.1093/aje/kwaa215

8. Dey T, Mukherjee A, Chakraborty S. A Practical Overview of Case-Control Studies in Clinical Practice. Chest. 2020;158(1 Suppl):S57–S64. doi:10.1016/j.chest.2020.03.033

9. Pearce N. What does the odds ratio estimate in a case-control study? Int J Epidemiol. 1993;22(6):1189–92. doi:10.1093/ije/22.6.1189

10. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol. 1982;116(3):547–53. doi:10.1093/oxfordjournals.aje.a113413

Epidemiology & Health Data Insights (eISSN 3080-8111)

Original Article. *Epidemiol Health Data Insights. 2025;1(1):ehdi003* https://doi.org/10.63946/ehdi/16386

Predictors of All-cause Mortality among Stroke Patients in Kazakhstan: A Retrospective Study Using Integrated AI Data Extraction

Artem Yermakov[1*], Kamila Akzholova[1*], Temirgali Aimyshev[1], Iliyar Arupzhanov[1], Gulnur Zhakhina[1], Abduzhappar Gaipov[1]

[1]*Department of Medicine, School of Medicine, Nazarbayev University, Astana, Kazakhstan*
* These authors have equally contributed to this article

Received: May 4, 2025
Accepted: May 16, 2025

Corresponding authors: Artem Yermakov & Kamila Akzholova

Email: artem.yermakov@nu.edu.kz & kamila.akzholova@nu.edu.kz

This work is licensed under a **Creative Commons Attribution 4.0 International License**

Abstract

Aim: To examine all-cause mortality predictors among lab indicators and drugs administered to stroke patients in Kazakhstan.

Methods: This retrospective study analyzed data from 272 patient records derived from the UNEHS database (2014-2019). The GPT-4o model was used to dissect the records and assist in the extraction of lab and medication data; other clinical and demographic data were derived from the previous study [6]. Statistical analyses included univariate, multivariate, and imputed multivariate logistic regressions (STATA version 16.1).

Results: In our cohort, deceased patients were older (66.1 vs. 58.6, $p < 0.0001$). Adjusted logistic regression revealed age (OR 1.04), hemoglobin (OR 0.97), and piracetam (OR 3.17) as independent predictors. After multiple imputation, Russian ethnicity (OR 2.79), PTI (OR 0.97), and dopamine (OR 5.23) became independent predictors, while piracetam lost its significance.

Conclusions: Findings suggest the importance of future research on stroke predictors in bigger cohorts to facilitate the introduction of innovations to stroke patients' care to potentially decrease the burden on the Kazakhstani population.

Keywords: Stroke; Retrospective cohort study; All-cause mortality; Predictors of mortality; AI-assisted data extraction; Regression analysis.

Introduction

Stroke has long been a major public health problem. In 2021, stroke was the third largest cause of death in the world [1]. Kazakhstan is no exception, and the third largest cause of death is stroke as well, taking 224.4 lives per 100,000 population in 2021 [2]. The American Stroke Association has defined stroke as "brain, spinal cord, or retinal cell death resulting from ischemia, based on pathological, imaging, or other objective evidence of cerebral, spinal cord, or retinal focal ischemic injury in a defined vascular distribution; or clinical evidence of cerebral, spinal cord, or retinal focal ischemic injury based on ≥24 hours of symptoms or until death, and excluding other causes" [3]. As indicated by the Global Burden of Disease (2024), in 2021, a stroke crude incidence was 11.95 million cases [4]. In the same year, the global prevalence of stroke was 93.82 million people, of which acute ischemic stroke accounted for 69.94 million, intracerebral hemorrhage for 16.60 million, and subarachnoid hemorrhage for 7.85 million cases. Looking at mortality, a total of 7.25 million stroke deaths were reported worldwide. Ischemic stroke caused 3.59 million deaths, intracerebral hemorrhage caused 3.31 million deaths, and subarachnoid hemorrhage caused 0.35 million deaths. [5]. In Kazakhstan, the incidence of stroke was valued at 1884 people per million population (PMP) [6].

Mortality from stroke is influenced by several factors, including stroke severity, type of the disease, and its complications [7, 8]. Early prediction of stroke outcome can help healthcare workers prioritize high-risk patients, optimize treatment regimens, and make better use of resources. This will improve the prognosis of patients, ultimately reducing disability and mortality rates among stroke patients. Different research works show that the utilization of laboratory testing results has great promise to be a reliable tool for that purpose [9-11].

Understanding the impact of drug treatment on stroke outcomes is essential for improving survival rates and guiding clinical decision-making. While pharmacological interventions, particularly thrombolytic therapy, antiplatelet agents, and anticoagulants, play a crucial role in stroke management, their effect on mortality in Kazakhstan remains underexplored.

The majority of the existing studies on mortality predictors are biased toward Western populations, therefore, it is important to contribute to global health equality by conducting similar studies on an underrepresented population of Kazakhstan. Unique demographic and genetic factors could influence the results and expand our understanding of the matter.

For that reason, this study aims to examine all-cause mortality predictors among lab indicators and drugs administered to stroke patients in Kazakhstan. We hypothesize that independent predictors of stroke mortality can be found in patients' laboratory and medication data.

Materials and Methods

Study Design and Population

This retrospective cohort study utilized hospital discharge records from the Unified National Electronic Health System (UNEHS) of Kazakhstan for patients admitted with stroke between 2014 and 2019 [12]. The original UNEHS records contained comprehensive patient data, including personal information, hospital and physician details, admission status, anamnesis, diagnosis, laboratory results, in-hospital drug treatment, and patient outcomes. Due to time constraints associated with the initial scope of this research as a Capstone project, detailed clinical data (laboratory indicators and drug treatment) were extracted for a randomly selected subset of 272 patient records. To obtain demographic characteristics for this cohort and supplement outcome data, these 272 records were subsequently merged with a larger dataset from a study by Zhakhina et al. (2022) and other UNEHS databases, using unique RPN IDs present in all records [6]. This approach also aimed to enrich the Zhakhina et al. data, which lacked detailed laboratory and medication information.

Ethical Considerations

The study received ethical approval from the Nazarbayev University Institutional Research Ethics Committee (NU-IREC 490/18112021). The requirement for informed consent was waived by the NU-IREC because the study was retrospective and utilized depersonalized data. All research activities were conducted in compliance with institutional requirements and relevant legal regulations.

Depersonalization Of Discharge Reports

Prior to any data extraction, all raw patient discharge reports underwent a depersonalization process. A custom-trained text processing model was developed and employed to remove all personal identifying information (PII), including names of patients, doctors, hospitals, and specific locations. This model achieved 100% accuracy on a manually labeled validation set of 100 records. The depersonalized reports were used for all subsequent steps.

Segmentation of Clinical Records

For the 272 selected patient records, clinical information within the depersonalized discharge reports was manually segmented by trained personnel. Records were divided into standardized sections: complaints, anamnesis, status, doctor's visits, laboratory tests, instrumental tests, drugs in hospital, procedures in hospital, diagnosis, prescription, recommendation, and outcome. This manual segmentation ensured accurate categorization for subsequent data extraction.

As part of a broader research initiative under the "Faculty Development Competitive Research (AI and Data Science) 2024-2026" Grant Program, an AI model (OpenAI GPT-4o-based) was also fine-tuned on approximately 2300 manually annotated records to perform automated segmentation (Figure 1). However, for the specific analysis presented in this paper, the data for the 272 patients were derived from the aforementioned manually segmented records to ensure the highest precision.

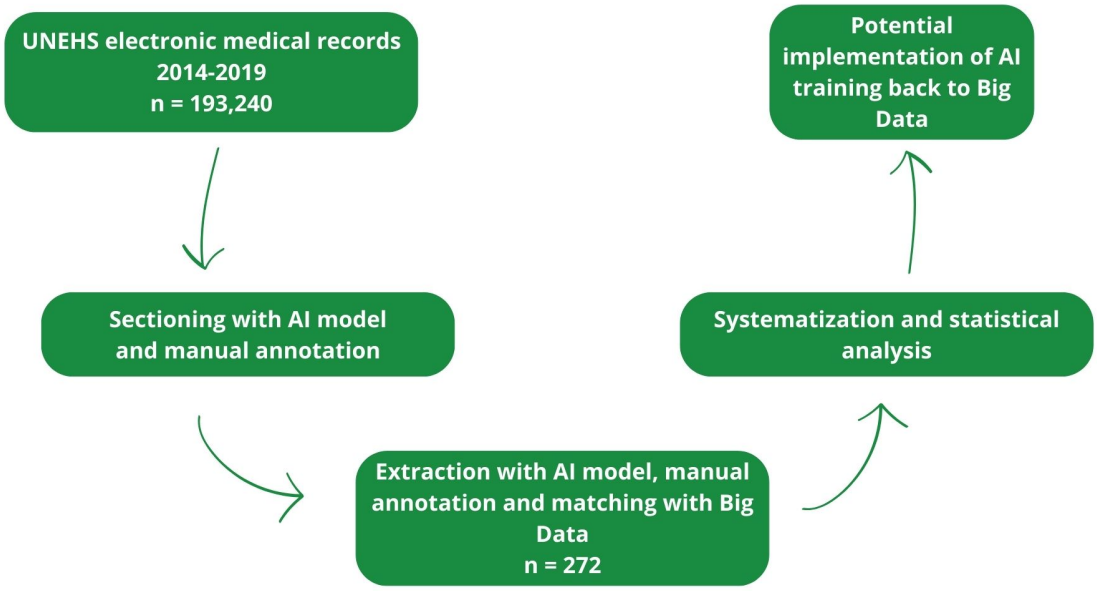

UNEHS electronic medical records
2014-2019
n = 193,240
Sectioning with AI model and manual annotation
Extraction with AI model, manual annotation and matching with Big Data
n = 272
Systematization and statistical analysis
Potential implementation of AI training back to Big Data

Figure 1. The data extraction process of electronic medical records.

Extraction of Laboratory and Medication Data

From the manually identified "drugs in hospital" and "laboratory tests" sections of the 272 records:

- Medication data: The OpenAI GPT-4o model was used for the initial extraction of drug names, dosages, and frequencies from the "drugs in hospital" segments. These AI-assisted extractions were subsequently manually reviewed, corrected, and validated utilizing a custom user interface (UI) depicted in Figure 2.
- Laboratory Data: Similarly, the GPT-4o model was used for the extraction of structured information from the "lab result" segments. All extracted laboratory values were then manually verified and corrected via a similar UI.

The resulting laboratory and medication data for the 272 patients reflect this dual, manually verified process.

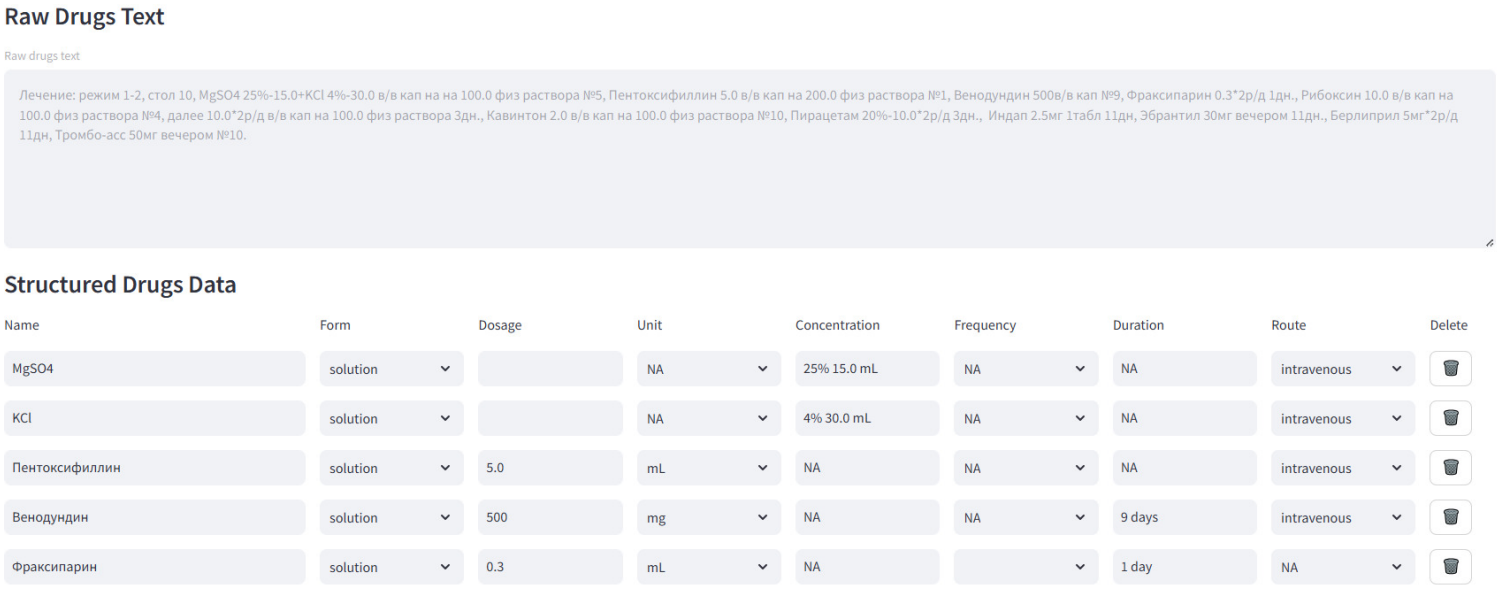


Figure 2. The interface of the training model for the extraction of drug treatment data.

Data Integration

Patient outcomes (e.g., mortality) for the 272-patient cohort were confirmed and supplemented by cross-referencing their RPN IDs with the database from the study by Zhakhina et al. (2022) and other databases within UNEHS [6]. The final analytical dataset combined the detailed clinical data from the 272 records with demographic and outcome data.

Definition of Variables

Demographic Characteristics.

Unique patient entries consisted of gender, date of birth (used to calculate age at admission), ethnicity, and type of admission. The date of death was recorded if applicable.

Medical Characteristics.

Medical variables included type of stroke (ischemic stroke [AIS], intracerebral hemorrhage [ICH], or subarachnoid hemorrhage [SAH]), follow-up period, length of hospital stay, and all-cause mortality. Acute kidney injury (AKI) was determined based on an increase in serum creatinine of more than 26 μmol/l between the first and second measurements during hospitalization, in accordance with National Kidney Foundation guidelines [13], where applicable.

Drug Treatment.

From the manually validated medication data, 173 unique drugs were identified. For this analysis, the use of 11 specific drugs frequently used in stroke treatment was coded as a binary variable (yes/no). These 11 drugs were: captopril, diazepam, dexamethasone, dopamine, magnesium sulphate, fraxiparin, actovegin, cerebrolysin, vinposan, piracetam, and strocitus. Their selection was based on the clinical protocol of the Ministry of Health and Social Development of the Republic of Kazakhstan (Protocol No. 7, July 13, 2016, code "I64 - stroke, not specified as hemorrhage or infarction").

Laboratory Tests.

The following laboratory indicators, extracted and verified as described above, were included in the analysis: creatinine, total serum bilirubin, fibrinogen, blood leukocyte count, hemoglobin, hematocrit, blood sugar, blood urea, erythrocyte sedimentation rate (ESR), potassium, sodium, blood protein, urea protein, alanine aminotransferase (ALT), aspartate transaminase (AST), prothrombin time (PTT), activated partial thromboplastin time (APTT), international normalized ratio (INR), and prothrombin index (PTI).

Statistical Analysis.

Descriptive data are depicted as percentages for categorical variables and as mean ± standard deviation (SD) in case of normal distribution or median (interquartile range—IQR) in other cases. All subjects were divided into survivors and deceased. The $\chi 2$ test was utilized for comparison of categorical variables, while continuous variables were compared using Student's t-tests or Mann–Whitney U tests, as appropriate. The association between demographics, laboratory data, medication data, and risk of all-cause mortality was examined using unadjusted (univariate) and adjusted (multivariate) logistic regression analyses. The following models were constructed that include confounders picked based on their significance level in unadjusted analysis and overall potential health effect: Model 1: age, gender, ethnicity, length of hospital stay; Model 2: added Dopamine, Fraxiparine, Piracetam, and Strocitus to Model 1; Model 3: added fibrinogen, hemoglobin, INR, PTI, blood sugar, and urea to Model 2. Subsequent logistic regression with multiple imputation was done as a sensitivity analysis to account for missing values (Supplementary Table 1).

All P values displayed are two-sided with the level of significance set at <0.05 for all analyses. All analyses were conducted using STATA 16.1 (College Station, TX).

Results

A total of 272 patients were included in the analysis. Among them, 126 (46.3%) died, and 146 (53.7%) survived during the follow-up period. The baseline demographic and clinical characteristics are shown in Table 1.

Table 1. General characteristics of admitted stroke patients.

Characteristic	Total (n = 272)	Survived (n = 146)	Deceased (n = 126)	p-value
Age, year (mean ± SD)	62.1±13.61	58.6±13.31	66.1±12.87	**<0.0001**
Gender, n (%)				0.380
Male	133 (48.90)	75 (51.37)	58 (46.03)	
Female	139 (51.10)	71 (48.63)	68 (53.97)	
Ethnicity, n (%)				**<0.0001**
Kazakh	146 (53.68)	91 (62.33)	55 (43.65)	
Russian	75 (27.57)	25 (17.12)	50 (39.68)	
Other	51 (18.75)	30 (20.55)	21 (16.67)	
Laboratory findings at admission				
Serum creatinine, μmol/l (median [IQR])	87.67 [68.3-113]	84 [62.31-105.7]	90 [75.9-120.6]	0.070
ALT, U (median [IQR])	16 [10-29.87]	16 [10-32.19]	14.9 [9.6-28]	0.453
aPTT, seconds (mean ± SD)	35.69±14.51	35.87±13.41	35.48±15.85	0.872
AST, U (median [IQR])	19.9 [10.8-31.8]	23.15 [11.4-32.2]	18.33 [10.8-29.9]	0.277
Total bilirubin, μmol/l (median [IQR])	13.68 [9.8-18.2]	13.9 [10-18]	13.2 [9.8-18.5]	0.582
Fibrinogen, g/l (mean ± SD)	3.81±1.65	3.60±1.42	4.09±1.9	0.077
Hemoglobin, g/l (mean ± SD)	136.19±21.57	139.04±19.46	132.58±23.61	**0.029**
Hematocrit, % (mean ± SD)	42.28±8.78	43.19±6.59	41.39±10.49	0.273
Leukocytes, *10^9/l (mean ± SD)	8.81±3.92	8.68±3.68	8.98±4.22	0.572
INR, (mean ± SD)	1.07±0.19	1.05±0.15	1.10±0.23	0.140
Potassium, mmol/l (mean ± SD)	4.07±0.86	4.13±0.92	3.99±0.80	0.423
Blood protein, g/l (mean ± SD)	67.95±10.81	68.48±10.26	67.28±10.13	0.472
PTI, % (mean ± SD)	91.87±10.43	94.29±8.46	89.22±11.73	**0.0049**
PTT, seconds (median [IQR])	15.5 [14-17.8]	15.2 [14-16.5]	16 [14.3-18.55]	0.118
Sodium, mmol/l (mean ± SD)	141.63±7.68	141.40±6.67	141.90±8.81	0.772
ESR, mm/h (median [IQR])	13 [6-23]	13 [5-23]	12.5 [6-23]	0.592
Blood sugar, mmol/l (median [IQR])	6.2 [5.3-7.85]	6.2 [5.3-7.9]	6.2 [5.3-7.7]	0.888
Urine protein, g/l (median [IQR])	0.03 [0-0.066]	0.033 [0-0.066]	0 [0-0.066]	0.939
Urea, mmol/l (median [IQR])	6.1 [4.6-7.6]	5.8 [4.4-6.9]	6.37 [5.2-8.82]	**0.0044**
Medications				
Captopril, n (%)	3 (1.10)	3 (2.05)	0 (0.00)	0.106
Diazepam, n (%)	3 (1.10)	0 (0.00)	3 (2.38)	0.061
Dexamethasone, n (%)	27 (9.93)	14 (9.59)	13 (10.32)	0.841
Dopamine, n (%)	14 (5.15)	3 (2.05)	11 (8.73)	**0.013**
Magnesium sulfate, n (%)	155 (56.99)	87 (59.59)	68 (53.97)	0.350
Fraxiparine (Nadroparin), n (%)	19 (6.99)	15 (10.27)	4 (3.17)	**0.022**
Actovegin, n (%)	141 (51.84)	74 (50.68)	67 (53.17)	0.682
Cerebrolysin, n (%)	32 (11.76)	19 (13.01)	13 (10.32)	0.491
Vinposan (Vinpocetine), n (%)	61 (22.43)	31 (21.23)	30 (23.81)	0.611
Piracetam, n (%)	133 (48.90)	64 (43.84)	69 (54.76)	0.072
Strocit (Citicoline), n (%)	62 (22.79)	39 (26.71)	23 (18.25)	0.097

Complication/outcome				
Length of hospital stay, days (median [IQR])	11.5 [9-15]	12 [10-16]	11 [8-14]	**0.042**
Follow up period, days (median [IQR])	2568 [330.5-2579]	2579[2572-2584]	187.5 [12-1086]	**<0.0001**
AKI, n (%)	12 (4.41)	5 (3.42)	7 (5.56)	0.393
Type of stroke, n (%)				0.965
AIS	118 (43.48)	63 (43.15)	55 (43.65)	
ICH	144 (52.94)	78 (53.42)	66 (52.38)	
SAH	10 (3.68)	5 (3.42)	5 (3.97)	
Type of death, n (%)				**<0.0001**
Early	44 (16.18)	0 (0.00)	44 (34.92)	
Late (>30 days)	82 (30.15)	0 (0.00)	82 (65.08)	

CKD, chronic kidney disease; ALT, alanine aminotransferase; APTT, activated partial thrombin time; AST, aspartate transaminase; INR, international normalized ratio; PTI, prothrombin index; PTT, prothrombin time; ESR, erythrocyte sedimentation rate; AKI, acute kidney injury; AIS, acute ischemic stroke; ICH, intracerebral hemorrhage; SAH, subarachnoid hemorrhage.

The mean age of the cohort was 62.1 ± 13.61 years, with non-survivors being older on average (66.1 ± 12.87) compared to survivors (58.6 ± 13.31). This difference is visualized in Supplementary figure 1. The gender distribution was 48.9% male and 51.1% female, with no significant difference between groups (p = 0.38). In terms of ethnicity, 53.68% were Kazakh, 27.57% Russian, and 18.75% other ethnicities, with a significant difference in mortality between these groups (p < 0.0001). Kazakhs were more likely to survive (62.33% survived), the opposite is true for Russians (33.3% survived), while among other ethnicities, there were a little bit more survivors than diseased patients (58.8% survived). For laboratory indicators, deceased patients had a lower mean hemoglobin (132.58 g/l vs. 139.04 g/l), a lower mean prothrombin index (89.22% vs. 94.29%), and a higher median urea level (6.37 mmol/l vs. 5.8 mmol/l) compared to survivors. Depiction of these differences can be found in Supplementary figures 2-3. Concerning drugs, deceased patients had a higher administration of dopamine (11 vs. 3) and a lower administration of fraxiparine (4 vs. 15).

Median length of hospitalization was 11.5 days, and survived patients stayed in the hospital for a longer time than non-survivors. The majority of patients (52.94%) were diagnosed with intracerebral hemorrhage, 43.48% of cases accounted for acute ischemic stroke, while subarachnoid hemorrhage accounted for the remaining 3.68%. Early death was less common among deceased patients, accounting for 34.92% of cases.

The univariate logistic regression was used to characterize associations between clinical-demographic parameters and death risk, and is summarized in Table 2. Each increase by 5 years of age resulted in a 26% higher risk of death (Odds Ratio – OR 1.26; 95% Confidence Interval – CI 1.14-1.40; p-value (p) < 0.0001). People of Russian ethnicity had almost 3-fold greater odds of death (OR 2.86; 95% CI 1.37-5.96; p = 0.005) compared to other nationalities (not Russian, nor Kazakh). Each incremental decrease by 10 g/l of hemoglobin is associated with a 15% higher risk of death (OR 1.15; 95% CI 1.01-1.31; p = 0.031). Each decrease of PTI by 10% increased odds of death by 37% (OR 1.37; 95% CI 1.01-1.84; p = 0.043). Urea showed a 2-fold increase in death odds with each elevation by 5 mmol/l (OR 2.01; 95% CI 1.25-3.23; p = 0.004). Administration of dopamine demonstrated a 4.5-fold increase in odds of death (OR 4.56; 95% CI 1.24-16.73; p = 0.022). Application of fraxiparine resulted in 71% lower risk of fatal outcome (OR 0.29; 95% CI 0.09-0.89; p = 0.030). Increase of follow-up period is connected to 9% lower odds of death (OR 0.91; 95% CI 0.86-0.96; p = 0.001).

Table 2. The association between demographics, medication, laboratory related parameters, and risk of all-cause mortality using univariate logistic regression models.

Characteristic	Number of observations / Number of events	Univariate OR (95% CI)	p-value
Age	272 / 126	1.05 (1.03-1.07)	**<0.0001**
Age +5	272 / 126	1.26 (1.14-1.40)	**<0.0001**
Gender (female ref.)			
Male	133 / 58	0.81 (0.50-1.30)	0.380
Ethnicity (other ref.)			
Kazakh	146 / 55	0.86 (0.45-1.66)	0.658
Russian	75 / 50	2.86 (1.37-5.96)	**0.005**
Serum creatinine	189 / 89	1.00 (1.00-1.01)	0.107
ALT	154 / 75	0.99 (0.97-1.01)	0.258
aPTT	146 / 66	0.998 (0.98-1.02)	0.871
AST	146 / 70	0.99 (0.98-1.01)	0.336
Total bilirubin	183 / 82	0.997 (0.97-1.02)	0.835
Fibrinogen	145 / 62	1.21 (0.97-1.49)	0.088
Hemoglobin	216 / 95	0.99 (0.97-1.00)	**0.031**
Hemoglobin -5	216 / 95	1.07 (1.01-1.15)	
Hemoglobin -10	216 / 95	1.15 (1.01-1.31)	
Hematocrit	115 / 58	0.98 (0.93-1.02)	0.280
Leukocytes	216 / 95	1.02 (0.95-1.09)	0.570
INR	148 / 62	3.80 (0.62-23.39)	0.150
Potassium	87 / 40	0.81 (0.49-1.35)	0.420
Blood protein	153 / 68	0.99 (0.99-1.02)	0.472
PTI	134 / 64	0.97 (0.94-1.00)	**0.043**
PTI -5	134 / 64	1.17 (1.01-1.36)	
PTI -10	134 / 64	1.37 (1.01-1.84)	
PTT	105 / 44	1.00 (0.97-1.03)	0.882
Sodium	81 / 37	1.01 (0.95-1.07)	0.769
ESR	181 / 82	1.01 (0.99-1.03)	0.335
Blood sugar	208 / 93	1.05 (0.98-1.11)	0.175
Urine protein	192 / 84	2.14 (0.57-7.95)	0.257
Urea	186 / 85	1.15 (1.05-1.26)	**0.004**
Urea +5	186 / 85	2.01 (1.25-3.23)	
Dexamethasone	272 / 126	1.09 (0.49-2.40)	0.841
Dopamine	272 / 126	4.56 (1.24-16.73)	**0.022**
Magnesium sulfate	272 / 126	0.80 (0.49-1.29)	0.351
Fraxiparine	272 / 126	0.29 (0.09-0.89)	**0.030**
Actovegin	272 / 126	1.10 (0.69-1.78)	0.682
Cerebrolysin	272 / 126	0.77 (0.36-1.63)	0.492
Vinposan	272 / 126	1.16 (0.66-2.05)	0.612

Piracetam	272 / 126	1.55 (0.96-2.51)	0.073
Strocit	272 / 126	0.61 (0.34-1.10)	0.099
Length of hospital stay	272 / 126	0.96 (0.93-1.00)	0.070
Follow-up period	272 / 126	0.91 (0.86-0.96)	**0.001**
AKI	272 / 126	1.66 (0.51-5.36)	0.398
Stroke type (AIS ref.)			
ICH	144 / 66	0.97 (0.60-1.58)	0.900
SAH	10 / 5	1.15 (0.32-4.17)	0.837

Captopril and diazepam were omitted due to collinearity. CKD, chronic kidney disease; ALT, alanine aminotransferase; APTT, activated partial thrombin time; AST, aspartate transaminase; INR, international normalized ratio; PTI, prothrombin index; PTT, prothrombin time; ESR, erythrocyte sedimentation rate; AKI, acute kidney injury; AIS, acute ischemic stroke; ICH, intracerebral hemorrhage; SAH, subarachnoid hemorrhage.

Adjustment using different models is depicted in Table 3. Age (OR 1.04; 95% CI 1.01-1.08; p = 0.023), hemoglobin level (OR 0.97; 95% CI 0.94-0.99; p = 0.008), and administration of piracetam (OR 3.17; 95% CI 1.20-8.41; p = 0.020). Dopamine showed collinearity.

Table 3. The association between demographics, medication, laboratory-related parameters, and risk of all-cause mortality using multivariate logistic regression models.

Covariates	Model 1 OR (95% CI)	p-value	Model 2 OR (95% CI)	p-value	Model 3 OR (95% CI)	p-value
Age	1.04 (1.02-1.06)	**<0.0001**	1.04 (1.01-1.08)	**0.017**	1.04 (1.01-1.08)	**0.023**
Gender (female ref.)	0.99 (0.59-1.65)	0.958	1.16 (0.50-2.73)	0.726	1.07 (0.42-2.74)	0.891
Ethnicity (other ref.)						
Kazakh	1.07 (0.54-2.10)	0.854	1.57 (0.54-4.55)	0.405	1.99 (0.61-6.44)	0.253
Russian	2.73 (1.28-5.81)	0.009	4.05 (1.18-13.86)	**0.026**	2.79 (0.74-10.52)	0.129
Hemoglobin	-	-	0.97 (0.95-0.99)	**0.011**	0.97 (0.94-0.99)	**0.008**
PTI	-	-	0.98 (0.95-1.01)	0.216	0.98 (0.94-1.01)	0.249
Urea	-	-	1.08 (0.97-1.20)	0.171	1.10 (0.96-1.27)	0.163
Dopamine	-	-	-		1	**-**
Fraxiparine	-	-	-		0.20 (0.03-1.28)	0.088
Piracetam	-	-	-		3.17 (1.20-8.41)	**0.020**
Strocit	-	-	-		0.38 (0.11-1.28)	0.119

Dopamine was omitted due to collinearity. INR, international normalized ratio; PTI, prothrombin index.

Table 4 illustrates the same adjusted models after multiple imputation was performed. Age (OR 1.04; 95% CI 1.02-1.07; p = 0.001) and hemoglobin level (OR 0.98; 95% CI 0.96-1.00; p = 0.045) remained independent predictors of mortality. Moreover, other parameters became significant predictors. Namely, Russian ethnicity (OR 2.84; 95% CI 1.19-6.79; p = 0.019) as well as PTI (OR 0.97; 95% CI 0.94-1.00; p = 0.043), and administration of dopamine (OR 5.23; 95% CI 1.06-25.74; p = 0.042). At the same time, piracetam was no longer significant in the model.

Table 4. The association between demographics, medication, laboratory-related parameters, and risk of all-cause mortality using multivariate logistic regression models with multiple imputation.

Covariates	Model 1 OR (95% CI)	p-value	Model 2 OR (95% CI)	p-value	Model 3 OR (95% CI)	p-value
Age	1.04 (1.02-1.06)	**<0.0001**	1.04 (1.02-1.07)	**0.001**	1.04 (1.02-1.07)	**0.001**
Gender (female ref.)	0.99 (0.59-1.65)	0.958	1.21 (0.66-2.21)	0.541	1.16 (0.63-2.14)	0.637
Ethnicity (other ref.)						
Kazakh	1.07 (0.54-2.10)	0.854	1.01 (0.47-2.15)	0.986	0.99 (0.45-2.15)	0.977
Russian	2.73 (1.28-5.81)	**0.009**	2.90 (1.23-6.83)	**0.015**	2.84 (1.19-6.79)	**0.019**
Hemoglobin	-	-	0.98 (0.97-1.00)	**0.047**	0.98 (0.96-1.00)	**0.045**
PTI	-	-	0.97 (0.94-1.00)	**0.041**	0.97 (0.94-1.00)	**0.043**
Urea	-	-	1.09 (0.99-1.20)	0.065	1.09 (0.98-1.21)	0.104
Dopamine	-	-	-		5.23 (1.06-25.74)	**0.042**
Fraxiparine	-	-	-		0.25 (0.07-1.01)	0.052
Piracetam	-	-	-		1.67 (0.94-2.97)	0.082
Strocit	-	-	-		0.83 (0.41-1.68)	0.600

INR, international normalized ratio; PTI, prothrombin index.

Discussion

Our results suggest that patients' age is connected to an elevation of odds of death in all regressions used. This result aligns with the findings of the source study [6]. Higher incidence of complications and higher mean number of chronic conditions in older people are possible explanations for this outcome [14].

Unadjusted logistic regression and adjusted model after imputation show that people of russian ethnicity have a substantially higher risk of death compared to Kazakhs and people of other ethnicities. Similar results are demonstrated in the source study [6], however odds ratio in our work is much higher. In our study, this discrepancy can potentially be attributed to age differences between these groups. After performing Student's t-test (Supplementary table 2), we found out that in this study Russian population was older on average than the Kazakh (66.9 vs. 58.8; t = 4.33; p = < 0.0001). Unfortunately, there are no appropriate studies we could find. Furthermore, our cohort has significantly more Kazakh population, and this ethnic disbalance is a limitation to this outcome.

We found that a higher blood urea level is a predictor of mortality in our unadjusted logistic regression. Previous studies mainly focused on blood urea nitrogen (BUN) or the BUN to creatinine ratio [9,10]. BUN is a component of urea that we have in our study, thus, these measurements are equivalent [15]. In case of our cohort, higher blood urea level is connected to higher odds of death, which aligns with findings of the aforementioned studies. This could be explained by the fact that high blood urea level is a crucial indicator of kidney damage. Therefore, patients with higher levels of urea in their blood were likely to have renal dysfunction, which would substantially increase the risk of death [16].

In previous research, it was found that higher admission levels of hemoglobin increased the odds of stroke patients' survival [11]. Our results support this claim, since lower levels of hemoglobin showed higher odds of fatality in all analyses. Anemia (low hemoglobin) was found to be associated with mortality in stroke patients in the research by Desai et al. (2023) [17]. On top of that, lower hemoglobin is an indicator of iron deficiency, which also contributes to higher odds of stroke mortality [18]. The possible explanations for this are a lowered ability of blood to supply organs, including the brain, with oxygen, which could reduce the chances of patients' survival during ischemic stroke [19].

Prothrombin index (PTI) is a mathematical ratio of reference thrombin time and a patient's thrombin time multiplied by 100% [20]. It is an indicator of coagulation, and a lower PTI value constitutes lower coagulation. This research suggests that lower PTI is a predictor of stroke mortality both in unadjusted and imputed

multivariate logistic regressions. This claim is backed up by previous research on anticoagulants' property to increase the odds of death in cerebral hemorrhage [21]. The majority of our cohort consisted of patients with intracerebral hemorrhage, and a lower ability of blood to coagulate would make this condition more severe due to higher volumes of blood entering the intracerebral space. This is a possible explanation of PTI's significance as a parameter of unadjusted regression in this study.

This research identified that administration of dopamine was a significant indicator of a higher risk of death in unadjusted logistic regression and adjusted regression after imputation, although the 95% confidence interval (CI) is very large. Dopamine is used as a peripheral vasostimulant. Low infusion rates (0.5-2 μg/kg/min) promote visceral vasodilation and can be used during renal failure to increase urinary flow. Intermediate infusion rates (2-10 μg/kg/min) elevate cardiac output via stimulation of electrical conductivity and myocardial contractility. Higher doses cause vasoconstriction and increased blood pressure [22]. We were unable to find high-quality studies regarding dopamine in stroke patients, however, the information we obtained suggests that dopamine is often used in critical care, where patients have lower chances of survival. This is the most probable explanation of the results in our study; however, it could also be due to chance, taking into consideration the large 95% CI.

Another parameter that was significant in the univariate regression is the administration of fraxiparine. Fraxiparine (nadroparin) is an anticoagulant sometimes used post-stroke for prophylaxis of deep vein thrombosis (DVT) [23]. Thus, fraxiparine was often administered to surviving patients to prevent complications, which could explain the lower odds of death.

Piracetam turned out to be a significant predictor of stroke mortality only in adjusted logistic regression but lost its significance after imputation. Manufacturers claim that this nootropic drug has neuroprotective properties and can ease cognitive recovery after stroke, although existing literature does not provide sufficient evidence regarding the effect of piracetam on stroke rehabilitation [24]. The large 95% CI in adjusted regression suggests that this result could appear due to chance.

This study is primarily limited by a small sample size, which has caused imbalances in ethnic, gender, and stroke type representation in our cohort. Furthermore, the manual evaluation of extracted data could have suffered from inconsistencies in parameter names and units of measurement. Another limitation arises from laboratory results, including a lot of missing values, although we addressed it using multiple imputation. Moreover, the source study by Zhakhina, et al. (2022) included various surgery procedures and several comorbidities, unfortunately, we were not able to match these parameters with our cohort due to the inability to quickly retrieve these data and time constraints [6]. Finally, as time-to-death data were available, we attempted to perform a survival analysis. However, after adjustment for relevant covariates, the results were not statistically significant and, therefore, were not included in the main text. The corresponding results are presented in Supplementary Tables 3–4, and the Kaplan–Meier survival curves are shown in Supplementary Figures 4–6.

Despite the aforementioned limitations, this is one of the few studies that examines post-stroke risk factors. In contrast, pre-stroke risk factors have been extensively studied. For example, O'Donnell et al. (2016) identified ten modifiable risk factors (such as hypertension, physical inactivity, poor diet, and smoking) that collectively account for approximately 90% of all strokes worldwide [25]. Incorporating clinical post-stroke factors has the potential to enhance the accuracy of risk assessment in stroke patients.

Conclusion

In conclusion, this research retrospectively assessed records of 272 stroke patients utilizing AI-assisted data extraction and data from the previous study [6] to understand how laboratory and medication variables are associated with mortality. Age and hemoglobin level were the only significant predictors of mortality in the adjusted logistic regression. PTI and Dopamine administration also became significant after multiple

imputation, although Dopamine was likely administered to critical patients. These findings have the potential to become important for stroke prognosis and risk assessment in the future. Future research with a bigger population is required to expand understanding of the subject.

Author Contributions: Conceptualization, A. Y., K. A., and A. G.; methodology, A. Y., K. A., T. A., I. A., and A. G.; validation, A. Y. and A. G.; formal analysis, A. Y., K. A., and A. G.; investigation, A.Y. and K. A.; resources, T. A. and I. A.; data curation, A. Y., K. A., I. A., and G. Z.; writing – original draft preparation, A. Y. and K. A.; writing – review and editing, A. Y., T. A, and I. A.; visualization, K. A., T. A., and I. A.; supervision, A. G.; project administration, A. Y., K. A., and A. G.; funding acquisition – not applicable. All authors have read and agreed to the published version of the manuscript.

Disclosures: There is no conflict of interest for all authors.

Acknowledgments: None.

Funding: This study was supported by grants from the Nazarbayev University Faculty Development Competitive Research Grant Program (AI and Data Science) 2024-2026 (Funder Project Reference: 201223FD2604, title: An AI-based approach for analyzing electronic medical records: Prediction of healthcare outcomes and drug demand). A.G. is a PI of the project.

Supplementary materials:

The Supplementary information includes tables:

- Supplementary Table 1. Number of missing values in each parameter;
- Supplementary Table 2. Age difference between Russian and Kazakh patients;
- Supplementary Table 3. Crude Cox regression analysis;
- Supplementary Table 4. Adjusted Cox regression.

As well as figures:

- Supplementary Figure 1. Age distribution among survived and deceased patients;
- Supplementary Figure 2. Hemoglobin levels among survived and deceased patients;
- Supplementary Figure 3. PTI distribution among survived and deceased patients;
- Supplementary Figure 4. Kaplan-Meier survival analysis among stroke patients adjusted for age category;
- Supplementary Figure 5. Kaplan-Meier survival analysis among stroke patients adjusted for gender;
- Supplementary Figure 6. Kaplan-Meier survival analysis among stroke patients adjusted for nationality.

The file can be accessed using: https://www.journalehdi.com/suppfile/725/Supplementary-information.docx

References

1. World Health Organization. The top 10 causes of death. 2024. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

2. World Health Organization. Kazakhstan data. 2023. https://data.who.int/countries/398

3. Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J. J. B., Culebras, A., Elkind, M. S. V., George, M. G., Hamdan, A. D., Higashida, R. T., Hoh, B. L., Janis, L. S., Kase, C. S., Kleindorfer, D. O., Lee, J.-M., Moseley, M. E., Peterson, E. D., Turan, T. N., Valderrama, A. L., & Vinters, H. V. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013;44(7):2064-2089. https://doi.org/10.1161/STR.0b013e318296aeca

4. GBD 2021 Stroke Risk Factor Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study

2021. The Lancet Neurology. 2024;23(10):973-1003. https://doi.org/10.1016/s1474-4422(24)00369-7

5. American Heart Association. 2025 Heart Disease & Stroke Statistical Update Fact Sheet Global Burden of Disease. American Heart Association. 2025. https://professional.heart.org/-/media/phd-files-2/science-news/2/2025-heart-and-stroke-stat-update/factsheets/2025-stats-update-fact-sheet-global-burden-of-disease.pdf

6. Gulnur Zhakhina, Bakhytbek Zhalmagambetov, Arnur Gusmanov, Yesbolat Sakko, Sauran Yerdessov, Elzar Matmusaeva, Aliya Imanova, Crape, B., Sarria-Santamera, A., & Abduzhappar Gaipov. Incidence and mortality rates of strokes in Kazakhstan in 2014–2019. Scientific Reports. 2022;12(1). https://doi.org/10.1038/s41598-022-20302-8

7. Nakibuuka, J., Sajatovic, M., Nankabirwa, J., Ssendikadiwa, C., Furlan, A. J., Katabira, E., Kayima, J., Kalema, N., Byakika-Tusiime, J., & Ddumba, E. Early mortality and functional outcome after acute stroke in Uganda: Prospective study with 30-day follow-up. SpringerPlus. 2015;4(450). https://doi.org/10.1186/s40064-015-1252-8

8. Shah, B., Bartaula, B., Adhikari, J., Neupane, H. S., Shah, B. P., & Poudel, G. (2017). Predictors of in-hospital mortality of acute ischemic stroke in the adult population. Journal of Neurosciences in Rural Practice. 2017;8(4):591–594. https://doi.org/10.4103/jnrp.jnrp_265_17

9. Schrock, J. W., Glasenapp, M., & Drogell, K. (2012). Elevated blood urea nitrogen/creatinine ratio is associated with poor outcome in patients with ischemic stroke. Clinical Neurology and Neurosurgery. 2012;114(7):881-884. https://doi.org/10.1016/j.clineuro.2012.01.031

10. You, S., Zheng, D., Zhong, C., Wang, X., Tang, W., Sheng, L., Zheng, C., Cao, Y., & Liu, C.-F. Prognostic Significance of Blood Urea Nitrogen in Acute Ischemic Stroke. Official Journal of the Japanese Circulation Society. 2018;82(2):572-578. https://doi.org/10.1253/circj.CJ-17-0485

11. Cohen, A. S., Siddiqui, K. A., Ali, S. F., Anderson, C. D., Silverman, S. B., & Schwamm, L. H. Abstract T P267: Admission Hemoglobin and Sodium Values are Important Factors Associated with Good Outcome After Acute Ischemic Stroke. Stroke. 2015;46(1). https://doi.org/10.1161/str.46.suppl_1.tp267

12. Arnur Gusmanov, Gulnur Zhakhina, Sauran Yerdessov, Yesbolat Sakko, Kamilla Mussina, Aidar Alimbayev, Dmitriy Syssoyev, Sarría-Santamera, A., & Abduzhappar Gaipov. Review of the research databases on population-based Registries of Unified electronic Healthcare system of Kazakhstan (UNEHS): Possibilities and limitations for epidemiological research and Real-World Evidence. International Journal of Medical Informatics. 2022;170:104950–104950. https://doi.org/10.1016/j.ijmedinf.2022.104950

13. National Kidney Foundation. Acute kidney injury (AKI). National Kidney Foundation. 2024. https://www.kidney.org/kidney-topics/acute-kidney-injury-aki

14. Chen, C.-M., Hsu, H.-C., Chang, C.-H., Lin, C.-H., Chen, K.-H., Hsieh, W.-C., & Chang, W.-M. Age-based prediction of incidence of complications during inpatient stroke rehabilitation: a retrospective longitudinal cohort study. BMC Geriatrics. 2014;14(1). https://doi.org/10.1186/1471-2318-14-41

15. Lab Tests Guide. Blood Urea VS BUN Definitions, Differences, and Similarities. Lab Tests Guide. 2023. https://www.labtestsguide.com/blood-urea-vs-bun

16. Tsagalis, G., Akrivos, T., Alevizaki, M., Manios, E., Stamatellopoulos, K., Laggouranis, A., & Vemmos, K. N. (2008). Renal dysfunction in acute stroke: an independent predictor of long-term all combined vascular events and overall mortality. Nephrology Dialysis Transplantation. 2008;24(1):194-200. https://doi.org/10.1093/ndt/gfn471

17. Desai, A., Oh, D., Rao, E. M., Sahoo, S., Mahajan, U. V., Labak, C. M., Rohit Mauria, Shah, V. S., Nguyen, Q., Herring, E., Elder, T., Stout, A., & Shammassian, B. H. Impact of anemia on

acute ischemic stroke outcomes: A systematic review of the literature. Impact of Anemia on Acute Ischemic Stroke Outcomes: A Systematic Review of the Literature. 2023;18(1):e0280025-e0280025. https://doi.org/10.1371/journal.pone.0280025

18. Gill, D., Monori, G., Tzoulaki, I., & Dehghan, A. Iron Status and Risk of Stroke. Stroke. 2018;49(12):2815-2821. https://doi.org/10.1161/strokeaha.118.022701

19. Kimberly, W. T., Wu, O., E. Murat Arsava, Garg, P., Ji, R., Vangel, M., Singhal, A. B., Ay, H., & Sorensen, A. G. Lower Hemoglobin Correlates with Larger Stroke Volumes in Acute Ischemic Stroke. Cerebrovascular Diseases Extra. 2011;1(1): 44–53. https://doi.org/10.1159/000328219

20. Sosudinfo. Protrombin, protrombinovuy indeks I Vremya (PTI, PTV) – ponyatie, rol, normy i otkloneniya [in Russian]. 2024. https://sosudinfo.ru/krov/protrombin/#h2_5%20

21. Morotti, A., & Goldstein, J. N. Anticoagulant-associated intracerebral hemorrhage. Brain Hemorrhages. 2020;1(1):89–94. https://doi.org/10.1016/j.hest.2020.01.001

22. Sonne, J., & Lopez-Ojeda, W. Dopamine. StatPearls Publishing. 2025. https://www.ncbi.nlm.nih.gov/books/NBK535451/

23. Hoh, B. L., Ko, N. U., Sepideh Amin-Hanjani, Sherry Hsiang-Yi Chou, Cruz-Flores, S., Dangayach, N. S., Derdeyn, C. P., Du, R., Hänggi, D., Hetts, S. W., Ifejika, N. L., Johnson, R., Keigher, K. M., Leslie-Mazwi, T. M., Lucke-Wold, B., Rabinstein, A. A., Robicsek, S. A., Stapleton, C. J., Suarez, J. I., & Tjoumakaris, S. I. 2023 Guideline for the Management of Patients With Aneurysmal Subarachnoid Hemorrhage: A Guideline From the American Heart Association/American Stroke Association. Stroke. 2023;54(7). https://doi.org/10.1161/str.0000000000000436

24. Wheble, P. C. R., Sena, E. S., & Macleod, M. R. A Systematic Review and Meta-Analysis of the Efficacy of Piracetam and Piracetam-Like Compounds in Experimental Stroke. Cerebrovascular Diseases. 2007;25(1-2):5–11. https://doi.org/10.1159/000111493

25. O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., Rao-Melacini, P., Zhang, X., Pais, P., Agapay, S., Lopez-Jaramillo, P., Damasceno, A., Langhorne, P., McQueen, M. J., Rosengren, A., Dehghan, M., Hankey, G. J., Dans, A. L., Elsayed, A., & Avezum, A. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet*. 2016;388:761–775. https://doi.org/10.1016/S0140-6736(16)30506-2